Multimedia Event Detection Task

The TRECVID 2010 Evaluation

Brian Antonishek, Jonathan Fiscus, Martial Michel, Paul Over
NIST

Stephanie Strassel, Amanda Morris
LDC

Motivation

- Current multimedia search technologies provide limited search capabilities from content directly extracted from the audio/visual signal and these approaches largely rely on human annotations

- MED addresses these limitations with a large collection of Internet videos, this domain presents many challenges
  - Variety of genres: Home video, interviews, tutorials, demonstrations, etc.
  - Variety of recording devices: Cell phone video, consumer video, professional equipment
  - Variety of cinematic effects: viewing angle, positioning, and motion
  - Variety of production: transitions (wipes, fades, etc.) and cinematography choices (time-lapse, filters, and lens)

Why a pilot study?

- Pilot aspects
  - Small data set
  - Small number of events
- Designed to answer certain questions to guide future evaluations
  - Is the task suitably challenging?
  - Which types of events can systems currently handle?
- Goals
  - Exercise the complete evaluation pipeline
  - Build the community

TRECVID MED
Multimedia Event Detection

- Task:
  - Given an event specified by a ***definition***, ***evidential description***, and ***illustrative examples***, detect the occurrence of the event within a multimedia clip
  - Identify each event observation by:
    - A ***binary decision*** on the detection score optimizing performance for the primary metric
    - A ***detection score*** indicating the system's confidence that the event occurred

The TRECVID MED 2010 Events

Test Event Definitions

Batting in a Run:

Within a single play during a baseball-type game, a batter hits a ball and one or more runners (possibly including the batter) scores a run.

Assembling a Shelter:

One or more people construct a temporary or semi-permanent shelter for humans that could provide protection from the elements.

Making a Cake:

One or more people make a cake.

The TRECVID MED 2010 Events

Event Name: Batting a run in

Definition:

Within a single play during a baseball-type game, a batter hits a ball and one or more runners (possibly including the batter) scores a run.

Evidential Description:

scene: outdoor or indoor ball fields (official or ad hoc), during the day or night

objects/people: baseball, bat, glove, crowd in background, fence, pitchers mound, bases, other players, officials

activities: pitching, swinging a bat, running, throwing a ball, cheering or clapping, making a call, crossing home plates

Exemplars:

http://www.flickr.com/photos/dustbowlballad/3283120050/

http://www.flickr.com/photos/amoney/3953671320/

http://www.flickr.com/photos/ricemaru/3500626769/

http://www.vimeo.com/5415112

Is this positive for “Batting a run in”?

The TRECVID MED 2010 Events

Event Name: Batting a run in

Definition:

Within a single play during a baseball-type game, a batter hits a ball and one or more runners (possibly including the batter) scores a run.

Evidential Description:

scene: outdoor or indoor ball fields (official or ad hoc), during the day or night

objects/people: baseball, bat, glove, crowd in background, fence, pitchers mound, bases, other players, officials

activities: pitching, swinging a bat, running, throwing a ball, cheering or clapping, making a call, crossing home plates

Exemplars:

http://www.flickr.com/photos/dustbowlballad/3283120050/

http://www.flickr.com/photos/amoney/3953671320/

http://www.flickr.com/photos/ricemaru/3500626769/

http://www.vimeo.com/5415112

Is this positive for “Batting a run in”?

The TRECVID MED 2010 Events

Event Name: Batting a run in

Definition:

Within a single play during a baseball-type game, a batter hits a ball and one or more runners (possibly including the batter) scores a run.

Evidential Description:

scene: outdoor or indoor ball fields (official or ad hoc), during the day or night

objects/people: baseball, bat, glove, crowd in background, fence, pitchers mound, bases, other players, officials

activities: pitching, swinging a bat, running, throwing a ball, cheering or clapping, making a call, crossing home plates

Exemplars:

http://www.flickr.com/photos/dustbowlballad/3283120050/

http://www.flickr.com/photos/amoney/3953671320/

http://www.flickr.com/photos/ricemaru/3500626769/

http://www.vimeo.com/5415112

Is this positive for “Batting a run in”?

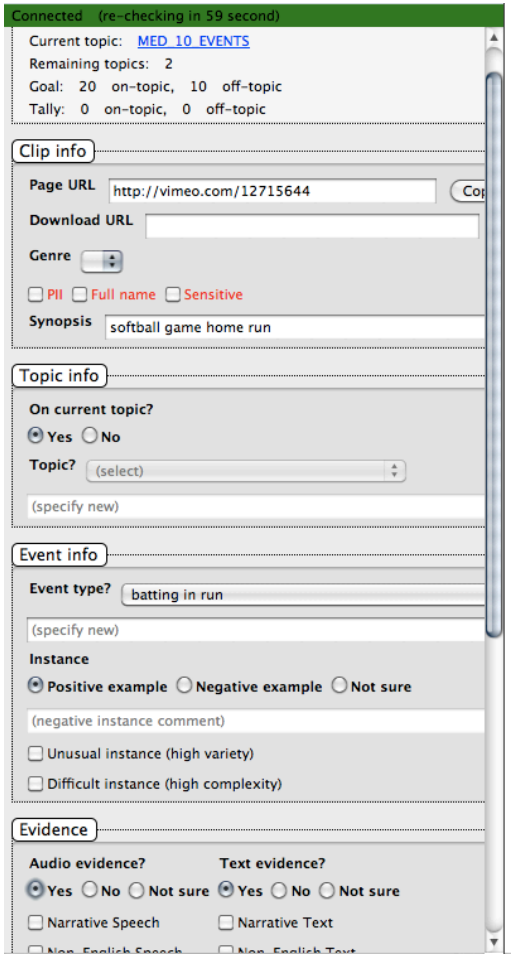
Data Collection & Annotation

- Team of 15 MED-10 data scouts at LDC
  - In-person training, regular team meetings, work remotely
- Custom GUI to search web for appropriate videos, then annotate their properties
- Two guiding annotation principles
  - **Sufficient Evidence Rule:** Video must contain sufficient evidence to decide that an event has occurred
    - ***Corollary****: Not necessary for video to contain every part of the event process to count as positive instance*
  - **Reasonable Viewer Rule:** If according to a reasonable interpretation of the video the event must have occurred, then the clip is a positive instance of that event

Annotation of Candidate Videos

- For each candidate video, scouts are required to
  - Watch clip in its entirety
  - Determine and verify the download URL
  - Screen for sensitive PII, objectionable content
  - Label event status (positive, negative, background)
- Each clip further annotated for
  - General topic category (sports, food, etc.)
  - Genre (home video, tutorial, amateur footage, etc.)
  - Brief synopsis
  - Optional: describe scene/setting, people/objects, activities
  - Optional: flag unusual or complex instances

AScout Screenshot



Quality Control and Validation

- All clips reviewed for licensing/IPR status
- After annotation, candidate clips are filtered to select those meeting corpus requirements
- Corpus clips undergo quality control review prior to distribution
  - All positive instances checked for annotation accuracy and completeness
  - Spot check on remaining clips based on combination of random and targeted clip selection

Data Processing for Distribution

- Automatic process downloads videos daily
- Downloaded videos processed to standardize data format and encoding
  - MPEG-4 format
  - h.264 video encoding
  - aac audio encoding
  - Original video resolution and audio/video bitrates retained
- Diagnostic information generated after processing
  - MD5 checksum
  - Duration

Source Data

Data Set	#Clips	#Hrs	Event Annotations						#Background
			Assembling a Shelter		Batting in a run		Making a Cake		
			#Pos.	#Neg.	#Pos.	#Neg.	#Pos.	#Neg.	
Training	1746	56	50	3	50	4	50	12	1577
Evaluation	1742	59	46	4	47	5	47	11	1582

Clip duration (both training and test)

	#Clips	Mean
All clips	3488	118s
Batting ev.	96	52s
Cake ev.	97	271s
Shelter ev.	97	158s

2010 Participants

7 Sites, 45 Submission Runs

		Number of Submissions		
		assembling_shelter	batting_in_run	making_cake
Center for Research and Technology, Hellas - Informatics and Telematics Institute	CERTH-ITI	9	9	9
Carnegie Mellon University	CMU	8	8	8
Columbia University / University of Central Florida	Columbia-UCF	6	6	6
IBM T. J. Watson Research Center / Columbia University	IBM-Columbia	10	10	10
KB Video Retrieval (Etter Solutions LLC)	KBVR	1	1	1
Mayachitra, Inc.	Mayachitra	2	2	2
Nikon Corporation	NIKON	9	9	9
Total Submissions per Event		45	45	45

Evaluation Protocol Synopsis

- Evaluation Plan
  http://www.nist.gov/itl/iad/mig/med.cfm
- Framework for Detection Evaluation (F4DE) Toolkit
  http://www.nist.gov/itl/iad/mig/tools.cfm
- Events are scored independently
- Evaluation process
  - Map system outputs onto the reference key
  - Error metric computation
  - Error Visualization

Metric computation

Missed Detection Probability (P_{Miss})

$$P_{Miss}(S,E_{i,}\Theta) = \frac{N_{Miss}(S,E_{i,}\Theta)}{N_{Target}(E_{i,})}$$

False Alarm Probability (P_{FA})

$$P_{FA}(S,E_{i,}\Theta) = \frac{N_{FA}(S,E_{i,}\Theta)}{N_{NonTarget}(E_{i,})}$$

Event Detection Constants

$Cost_{Miss} = 80$

$Cost_{FA} = 1$

$P_{Target} = 0.001$

$N_{Miss}(S,E_{i,}\Theta)$ = *number of missed detections for system S, event E_i at decision score Θ*

$N_{Target}(E_{i,})$ = *number of clips containing event instances for event E_i*

$N_{NonTarget}(E_{i,})$ = *number of clips that do not contain event instances for event E_i*

$N_{FA}(S,E_{i,}\Theta)$ = *number of false alarms for E_i at decision score Θ*

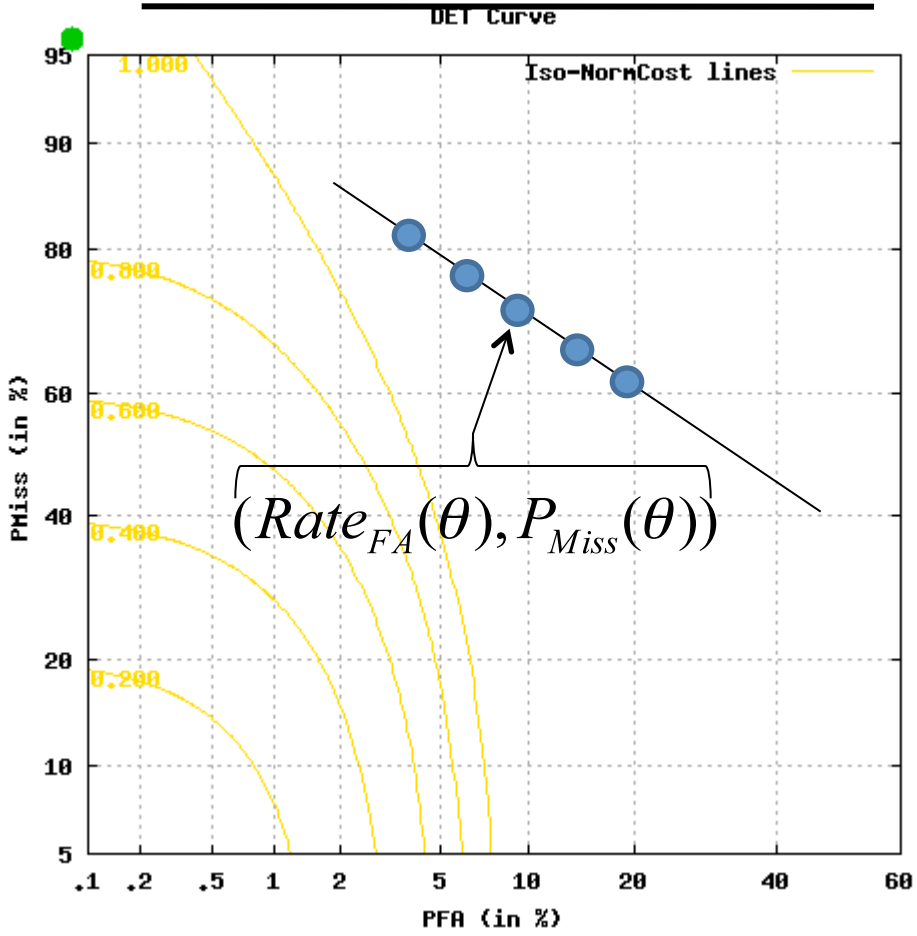
Normalized Detection Cost of a system (NDC)

$$NDC(S,E) = \frac{Cost_{Miss} * P_{Miss}(S,E) * P_{Target} + Cost_{FA} * P_{FA}(S,E) * (1 - P_{FA}(S,E))}{MINIMUM(Cost_{Miss} * P_{Target}, Cost_{FA} * (1 - P_{Target}))}$$

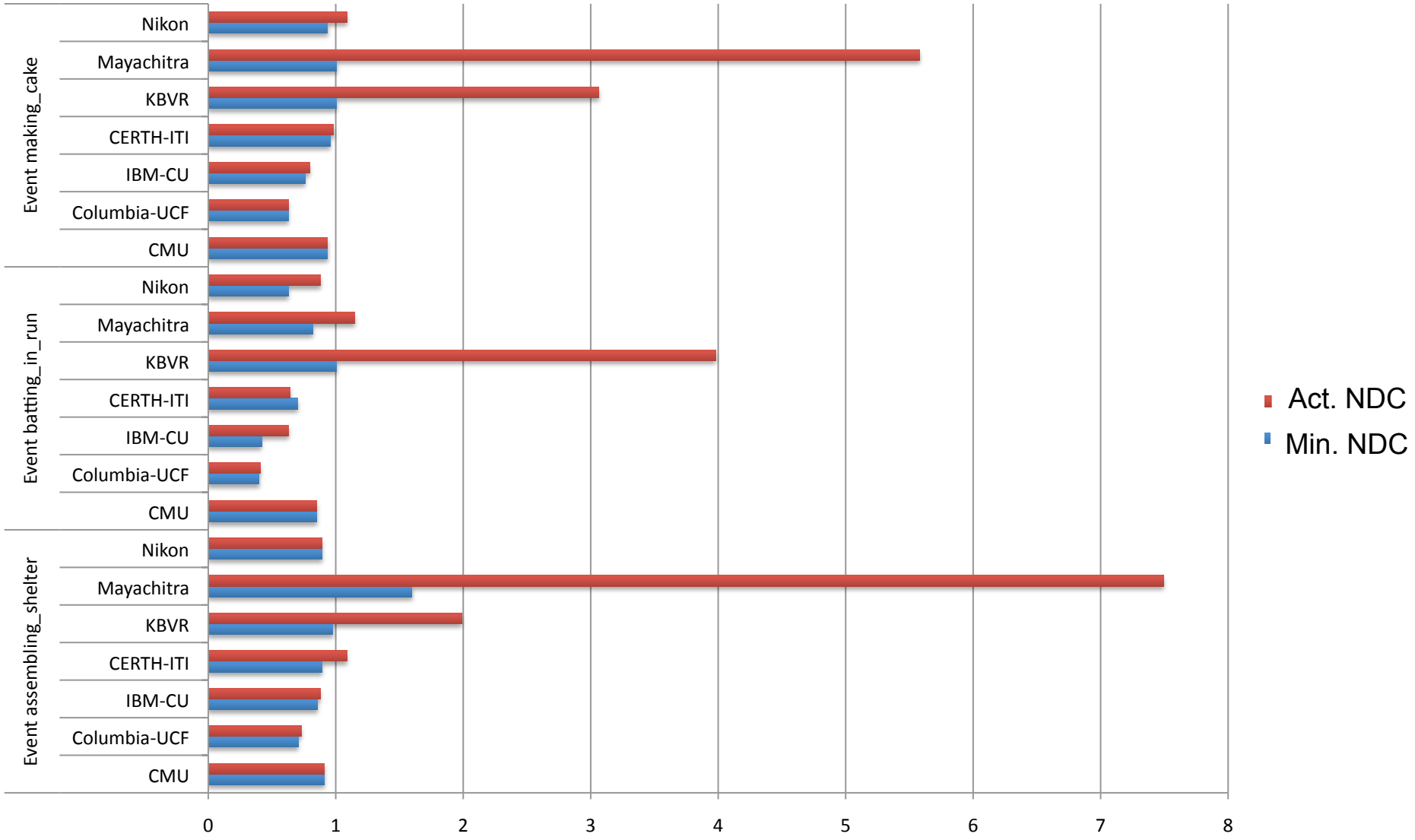
Decision Error Tradeoff (DET) Curves

$Prob_{Miss}$ vs. $Rate_{FA}$

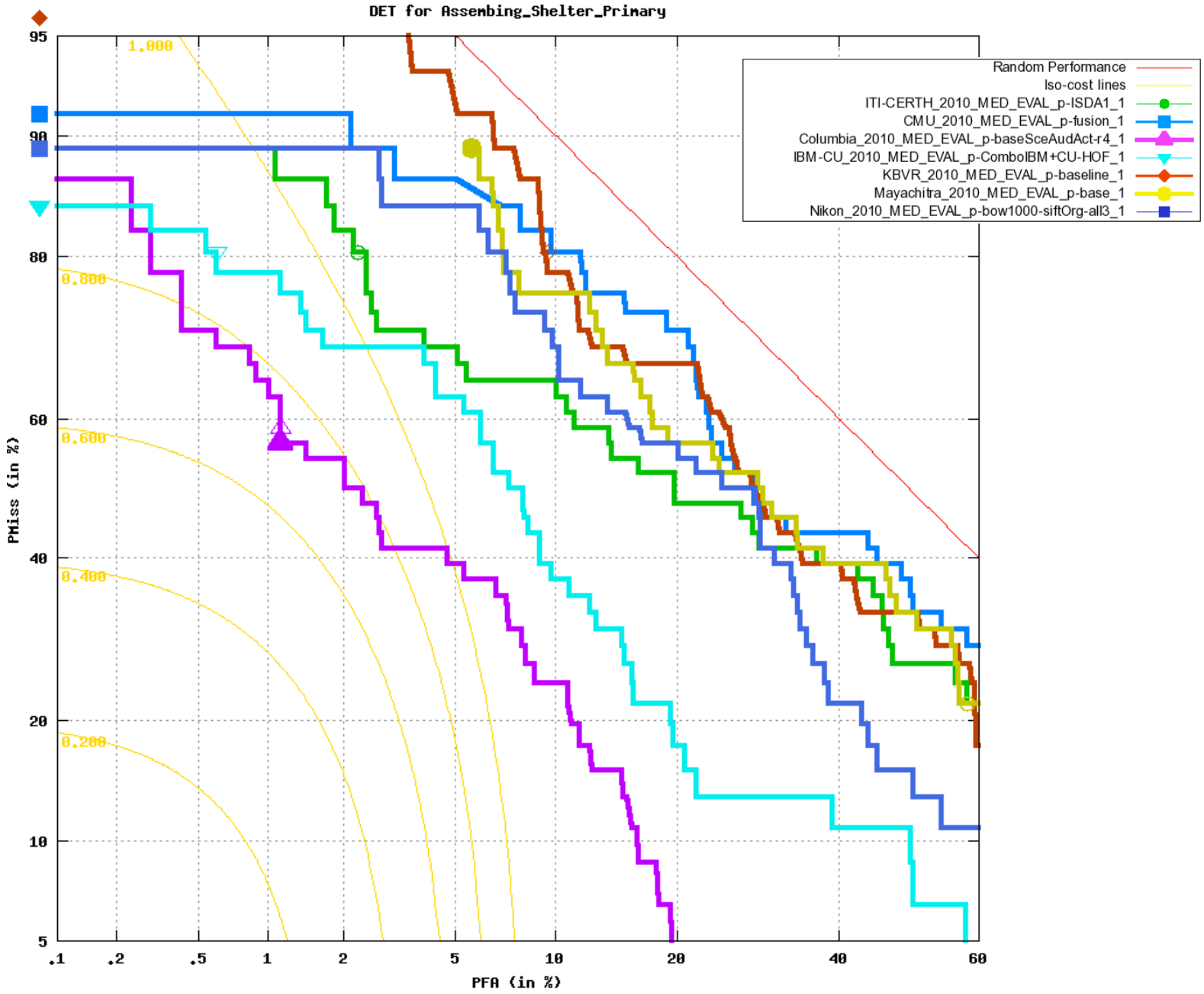
Compute $Rate_{FA}$ and P_{Miss} for all Θ



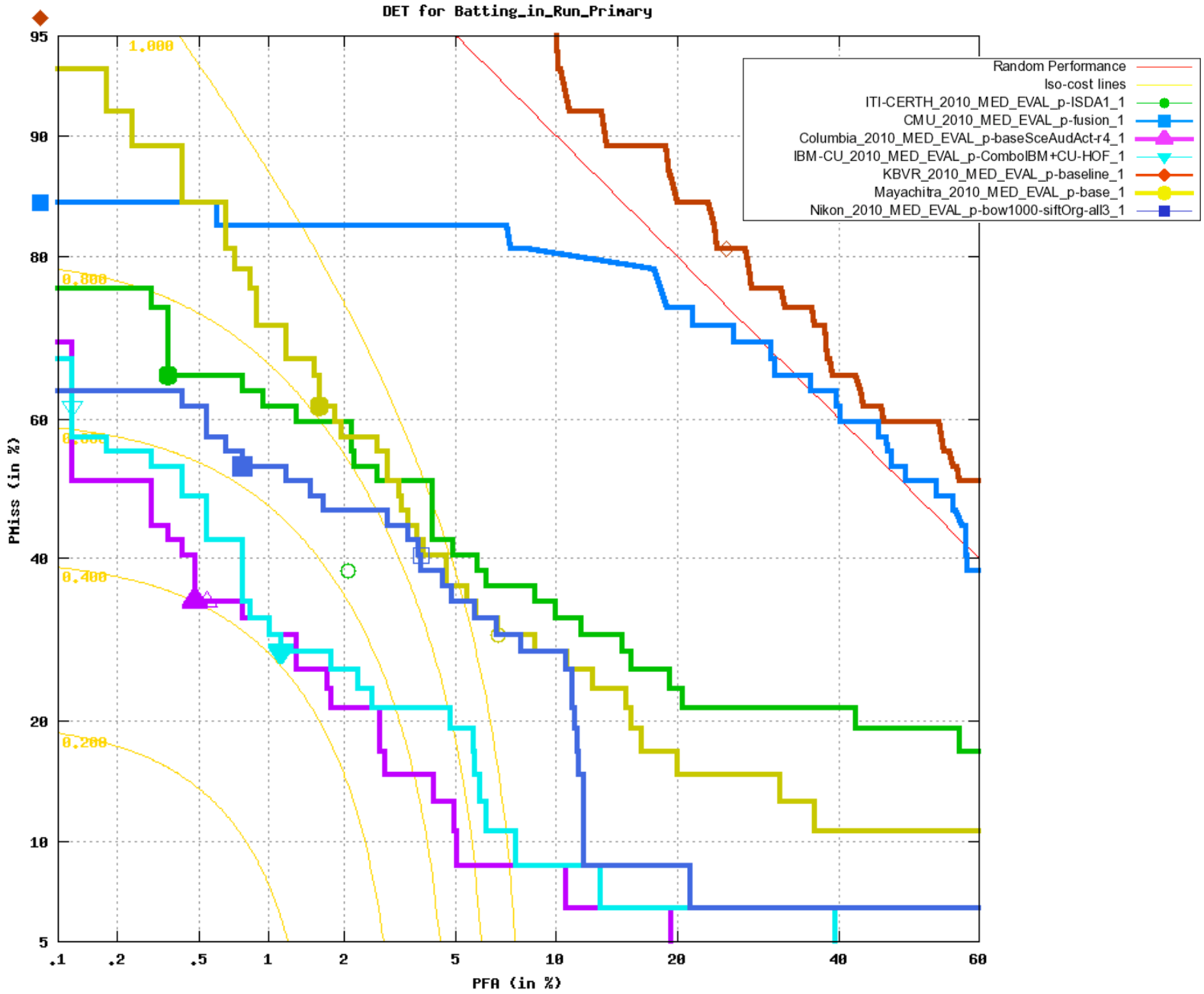
2010 Minimum and Actual Normalized Detection Cost (NDC)



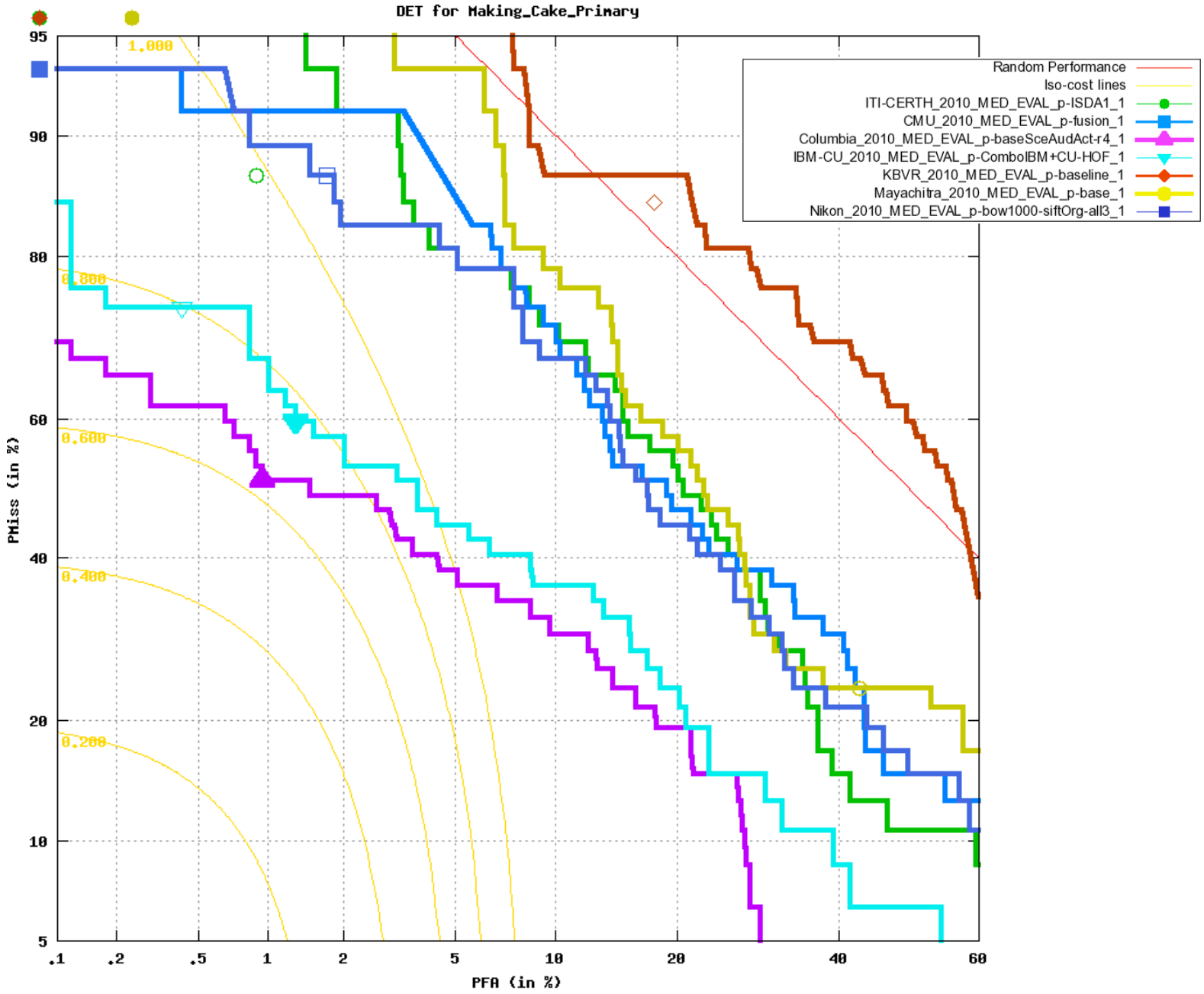
Assembling a Shelter (Primary systems)



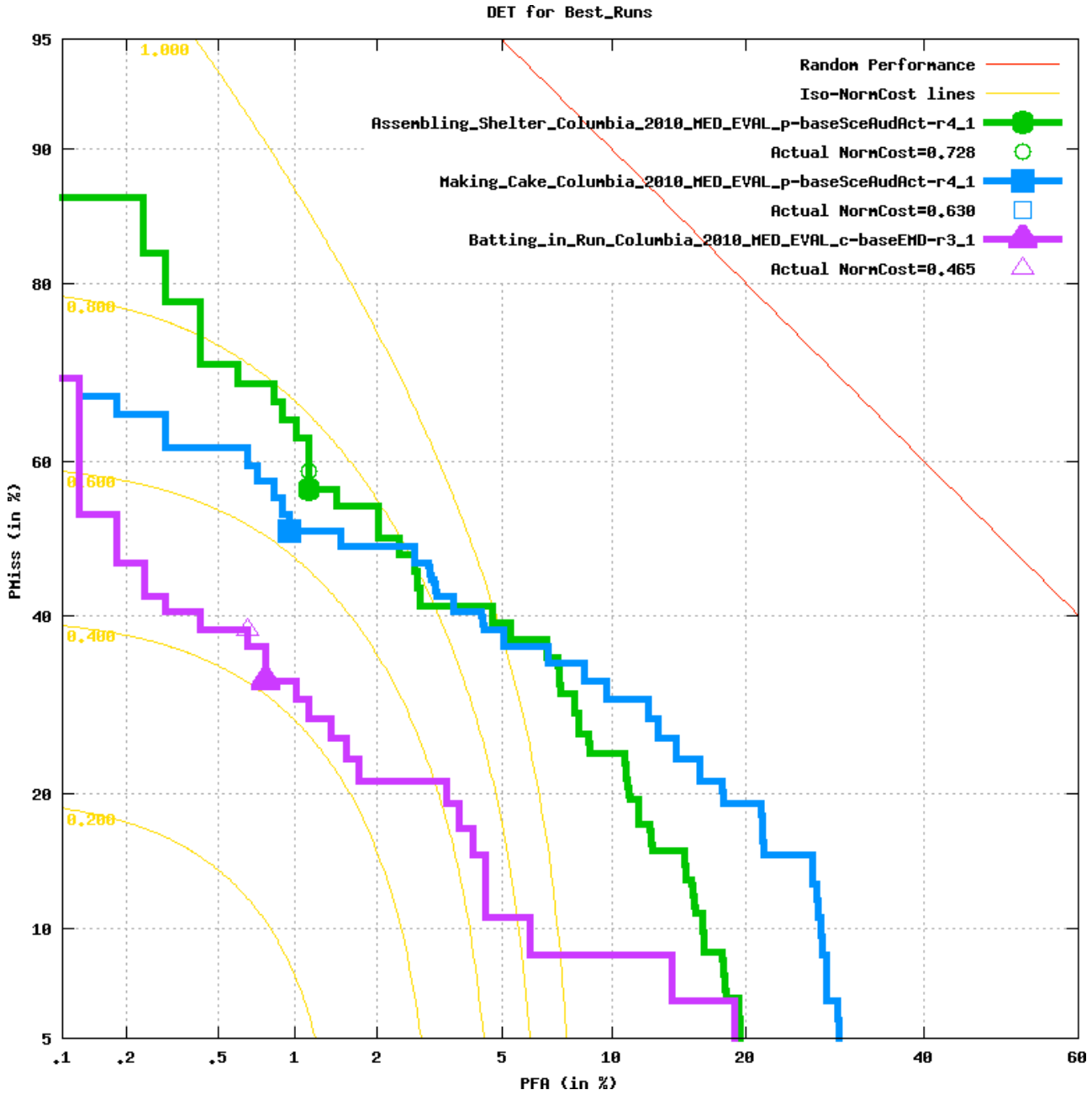
Batting in a Run (Primary systems)



Making a Cake (Primary systems)



“Best” run for each event



Conclusions and Lessons Learned

- Successful pilot evaluation
  - First use of the HAVIC corpus
  - Developed an event definition, evaluation task, performance metrics, and evaluation tools
- Surprising pilot results
  - Technology demonstrated the capability of detecting clips containing specified events.
- Analysis has just begun
  - Adjudication experiments (purify the references)
  - Measuring the impact negative event instances
- Next year?
  - More events and larger data sets will present greater challenges to the systems

Questions?