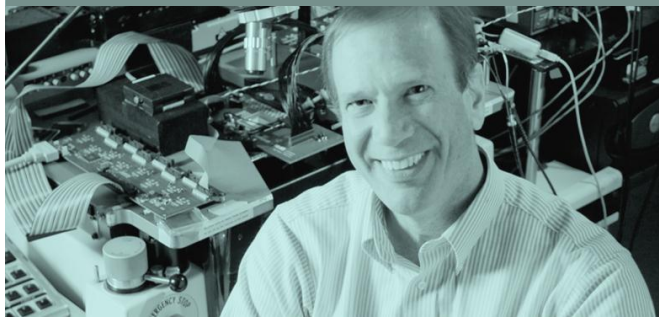
SRI International
SARNOFF



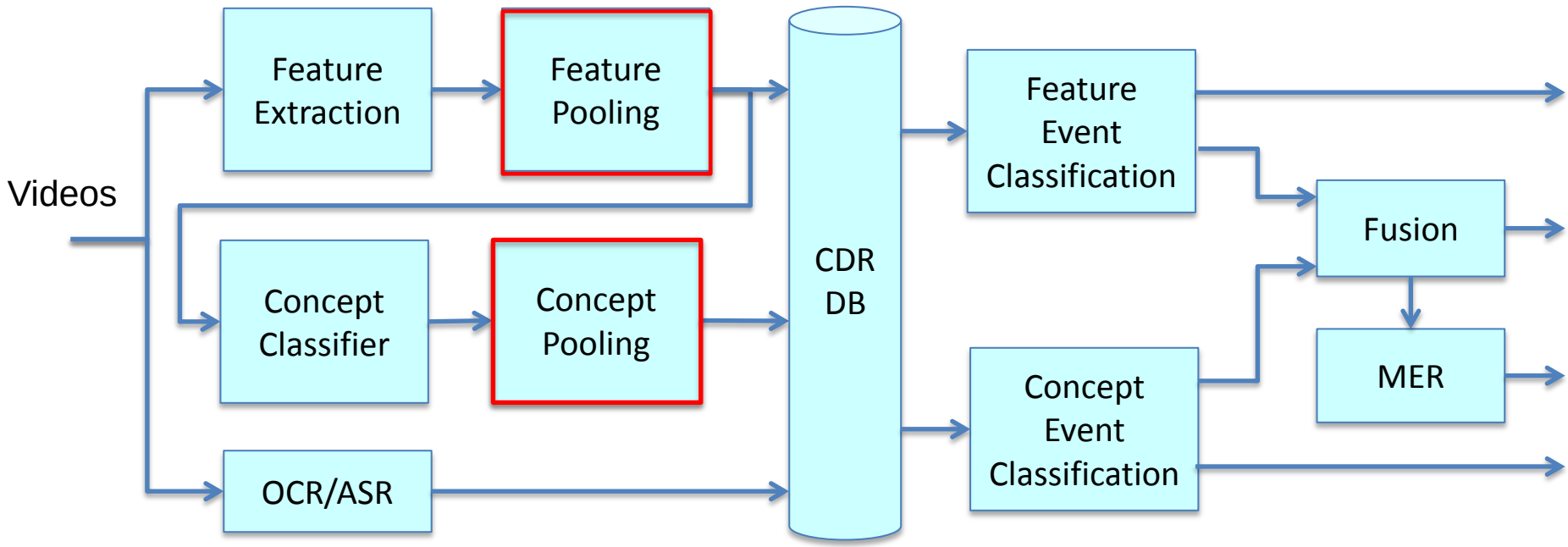
AURORA: Multimedia Event Detection (MED'12)

Team: SRI Sarnoff, CMU, Cycorp, ICSI, UCF and UMass,

Vision Technology Center, SRI International Sarnoff

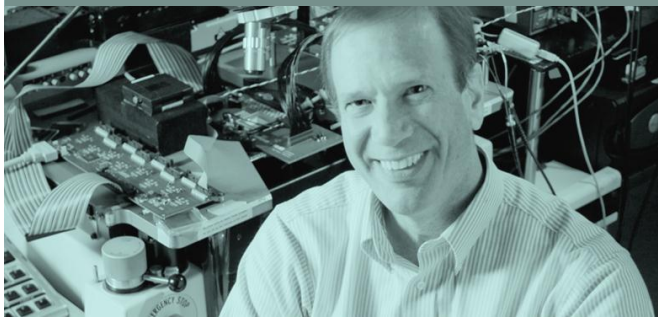
SRI International SARNOFF · Carnegie Mellon University · cycorp · ICSI International Computer Science Institute · Computer Vision Lab

AURORA MED'12 Processing Flow



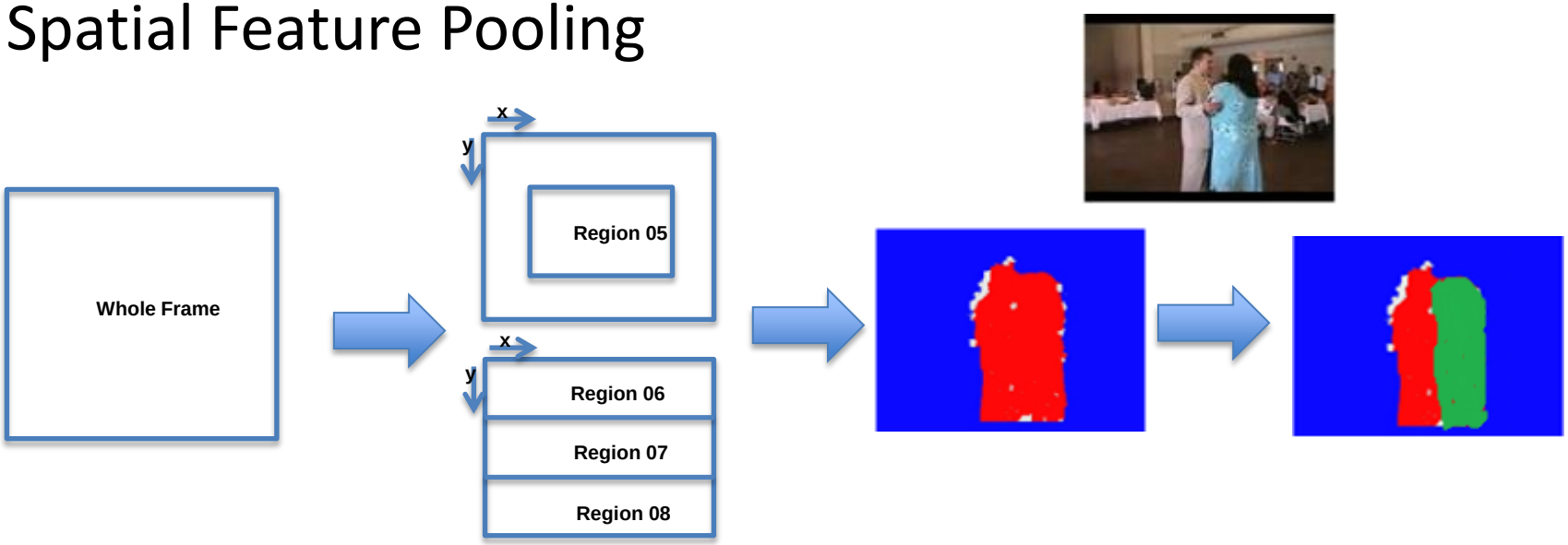
- (Low-Level) Feature Extraction: Dense Trajectory Feature (DTF: DTF_HOG and DTF_MBH), STIP, Sparse SIFT, Dense SIFT, Color SIFT, Motion SIFT , Transformed Color Histogram (TCH), MFCC
- Event representation: Bag of Visual Words
- Classification: SVM Classifier with various kernel

SRI International
SARNOFF



Low-Level Feature Pooling

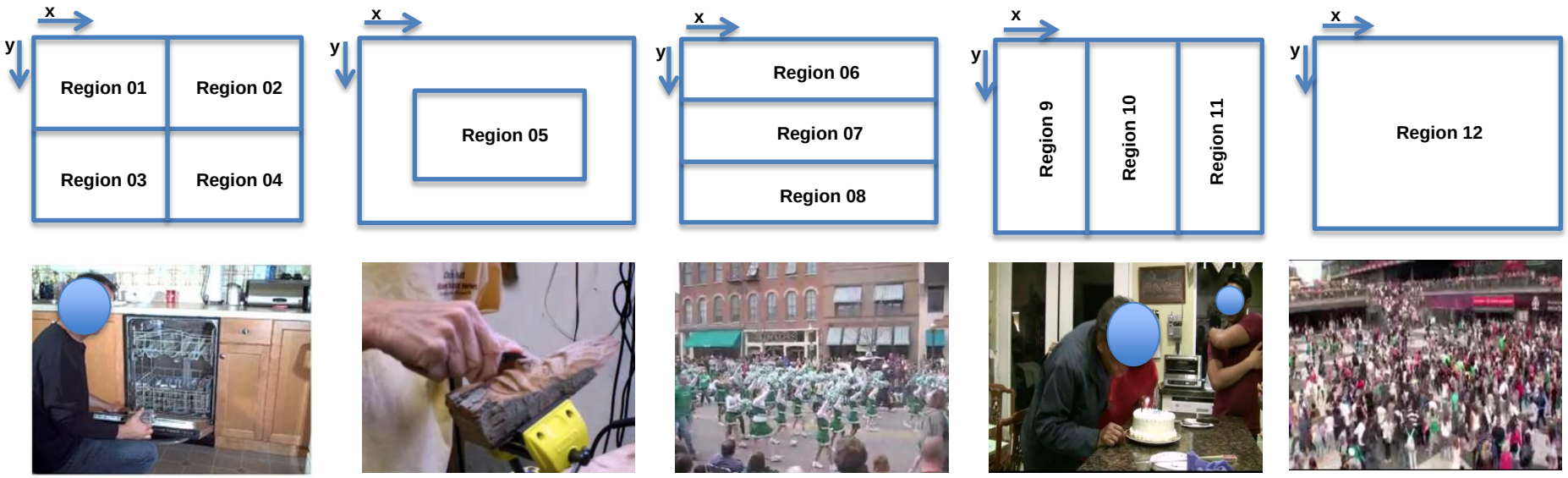
Spatial Feature Pooling



Whole Frame **Fixed Patterns** **Motion Layer** **Object**

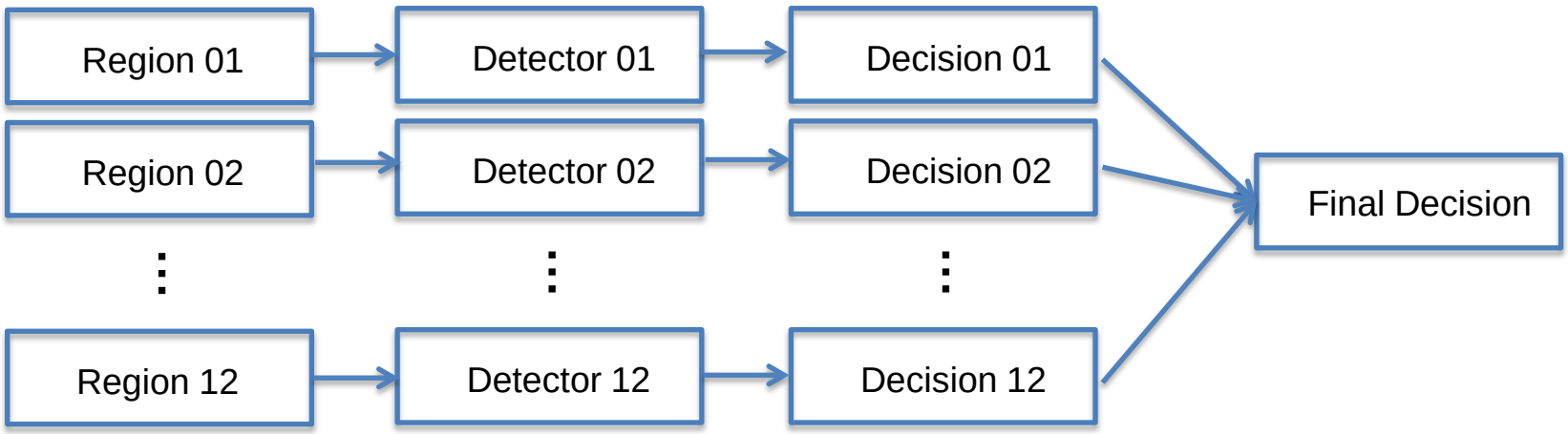
- From frame level feature extraction to localized feature extraction
  - Frame level: aggregate features from the whole frame
  - Fixed spatial patterns: aggregate features from pre-defined patterns
  - Motion layer based: aggregate features using each motion layers
  - Object based: aggregate features using object masks

Feature Pooling Using Fixed Spatial Patterns



- Objective
  - Limitation: Features aggregated from a whole frame contains more irrelevant data of an event
  - Goal: Extract event relevant information by pooling features from different parts of a frame
- Spatial pooling using fixed patterns
  - Aggregate features over a set of pre-defined regions as shown above
  - Implicitly encodes location information with visual-words for better matching
  - Fixed patterns are easy and fast to computer

Concept and Event Detection Using Spatial Pooling



- Concatenating region-based BOW models into a long vector for one video is often infeasible and inefficient for large codebook.
- Our solution
  - Treat each region as an individual information source, and train/recognize concepts/events at region level
  - Fuse the decisions from all region sources

Event Detection Using Spatial Pooling

Waypoint Experiment

- Single Feature: DTF_HOG
- Train on MED11 EC + DEVT, and test on MED11 DEVO
- MD@6%FA: about 2% ~ 11% improvement

	UL	UR	LL	LR	Ctr	Up	Md	Lw	Lf	Md	Rt	Full	All	
	R01	R02	R03	R04	R05	R06	R07	R08	R09	R10	R11	R12	R01-11	Imp.
E06 Birthday_party	40.7	44.8	44.2	39.0	37.2	46.5	37.2	41.9	43.0	39.5	39.0	31.4	20.9	10.5
E07 Changing_a_vehicle_tire	43.4	49.6	31.9	34.5	36.3	50.4	32.7	41.6	43.4	36.3	46.0	33.6	23.0	10.6
E08 Flash_mob_gathering	18.5	17.0	13.3	11.9	11.9	21.5	11.1	14.1	13.3	13.3	11.9	11.1	8.9	2.2
E09 Getting_a_vehicle_unstuck	36.1	33.7	34.9	30.1	27.7	42.2	31.3	30.1	39.8	28.9	33.7	27.7	21.7	6.0
E10 Grooming_an_animal	49.4	44.4	53.1	51.9	45.7	48.2	37.0	54.3	56.8	45.7	53.1	39.5	35.8	3.7
E11 Making_a_sandwich	57.7	67.2	48.9	54.7	54.0	55.5	47.5	56.9	54.7	51.8	58.4	49.6	38.7	10.9
E12 Parade	34.2	33.7	30.5	33.2	25.7	38.5	24.1	32.6	28.3	27.3	26.2	27.8	16.0	11.8
E13 Parkour	27.5	28.4	14.7	23.5	14.7	30.4	15.7	28.4	20.6	16.7	22.6	16.7	14.7	2.0
E14 Repairing_an_appliance	33.0	33.0	34.1	34.1	28.4	29.6	35.2	36.4	30.7	31.8	35.2	28.4	23.9	4.5
E15 sewing_project	52.4	52.4	47.6	52.4	47.6	56.1	40.2	52.4	57.3	45.1	54.9	43.9	37.8	6.1

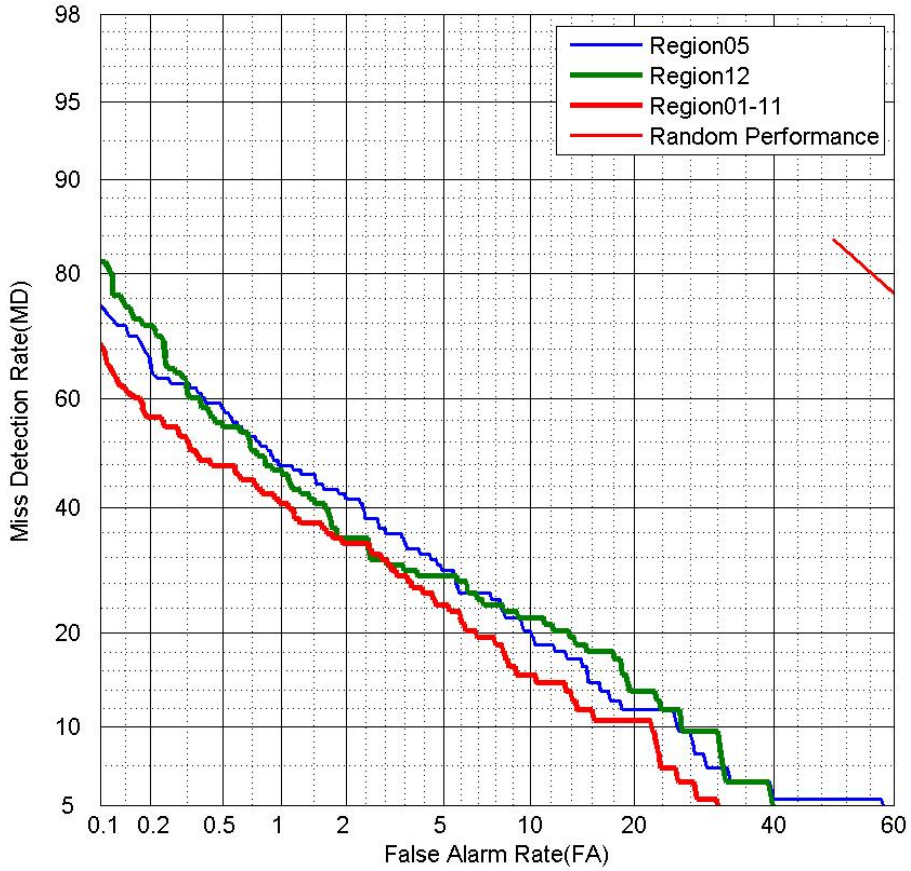
Improving Event Detection Using Spatial Pooling

- Features: Fusion of DTF_HOG, DTF_MBH, STIP, and SIFT
- Train on MED11 EC + DEVT, and test on MED11 DEVO
- MD@6%FA: about 1.1% ~ 6.2% improvement

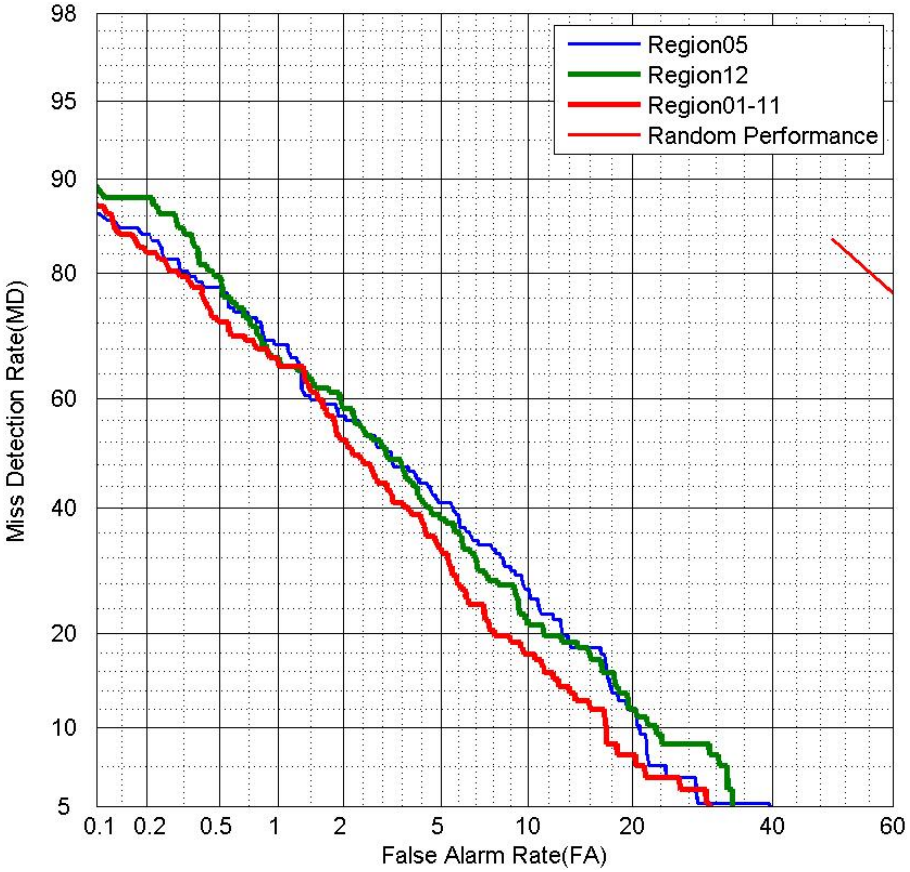
	UL	UR	LL	LR	Ctr	Up	Md	Lw	Lf	Md	Rt	Full	All	
(%)	R01	R02	R03	R04	R05	R06	R07	R08	R09	R10	R11	R12	R01-11	Diff
E06 Birthday_party	34.3	40.1	29.7	29.7	27.9	39.5	28.5	34.9	31.4	26.7	31.4	22.1	20.9	**1.2**
E07 Changing_a_vehicle_tire	28.3	34.5	32.7	36.3	25.7	32.7	25.7	39.8	32.7	26.6	34.5	27.4	21.2	**6.2**
E08 Flash_mob_gathering	10.4	11.1	8.1	6.7	6.7	13.3	7.4	7.4	7.4	7.4	7.4	7.4	5.9	**1.5**
E09 Getting_a_vehicle_unstuck	36.1	26.5	27.7	22.9	22.9	27.7	24.1	33.7	30.1	26.5	26.5	21.7	18.1	**3.6**
E10 Grooming_an_animal	43.2	37	50.6	48.2	28.4	44.4	30.9	42	46.9	35.8	53.1	33.3	30.9	**2.4**
E11 Making_a_sandwich	41.6	45.3	39.4	40.9	36.5	40.2	34.3	42.3	42.3	37.2	43.8	33.6	26.3	**2.3**
E12 Parade	19.3	18.7	20.9	19.3	11.8	24.6	15	19.8	16	13.9	11.8	11.8	10.7	**1.1**
E13 Parkour	10.8	11.8	9.8	11.8	5.9	15.7	5.9	13.7	11.8	4.9	9.8	5.9	3.9	**2**
E14 Repairing_an_appliance	25	23.9	27.3	22.7	20.5	21.6	23.9	28.4	29.6	26.1	26.1	19.3	18.2	**1.1**
E15 Sewing_project	35.4	40.2	34.2	37.8	34.2	36.6	36.6	35.4	43.9	34.2	37.8	30.5	29.3	**1.2**

DET Curves for Some Events

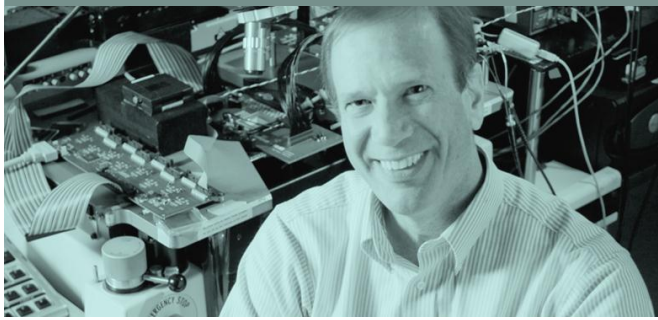
E07 Changing a Vehicle Tire



E11 Making a Sandwich



SRI International
SARNOFF



Semantic Concept (Pooling) Feature

Semantic Concepts

- Part I: MED11 and MED12 data related action concepts
  - Atomic and localized motion and appearance patterns
  - Event is usually strongly associated with some specific actions
- UCF action concepts
  - 93 concepts have been selected for MED12 evelation
  - Usually they are general concepts

Animal approaching	Person bending	Person climbing	Person eating	Person kissing
Person blowing candles	Animal eating	Person cutting	Person falling	Person laughing
Open door	Person_biking	Person_cleaning	Person_driving	Person_crying

- Sarnoff action concepts
  - 92 concepts have been selected for MED12 elevation
  - Focus on concepts more specific to an event

Standing on top of bike	Running next to dog	Delivering a speech	Removing debris
scrubbing appliance by a brush	Pouring melted metal	Breaking through tape	Measuring length
Moving along a rock face	Hooking rope to harness	Giving dogs treats	Cutting metal

Semantic Concepts

- Part II: Third-Party action concepts
  - UCF50: includes many professional sport actions, such as:

Horse Riding	Jumping Jack	Playing Violin	Rock Climbing Indoor	Diving
Baseball Pitch	High Jump	Skate Boarding	Bench Press	Golf Swing

  - Brown HMDB 51: daily life actions, such as:

Brush Hair	Kick Ball	Smoke	Punch	Drink	Eat	Climb Stairs	Dribble
Chew	Catch	Somersault	Clap	Sit	Smile	Shake Hands	Fencing

  - Some are relevant to MED 11 & 12 events, and others may be relevant to background events

Semantic Concepts

- PART III: SIN dataset (CMU)
  - TRECVID Semantic Indexing task
  - 346 generic concepts including object, scenes, and actions
  - All defined by static images (key frames, no motions)
  - **There is no overlap between SIN data set and ALLADIN event dataset**

River	Snow	Streets	Suburban	Dining_Room	Fields	Kitchen	Mountain
Truck	Vehicle	Cattle	Table	Body_Parts	Clouds	Dogs	Canoe
Door_Opening	Throwing	Skating	Walking	Exiting_Car	Standing	Cheering	Dancing
Rowboat	Handshaking	Road	Sky	People_Marching	Singing	Animal	Horse

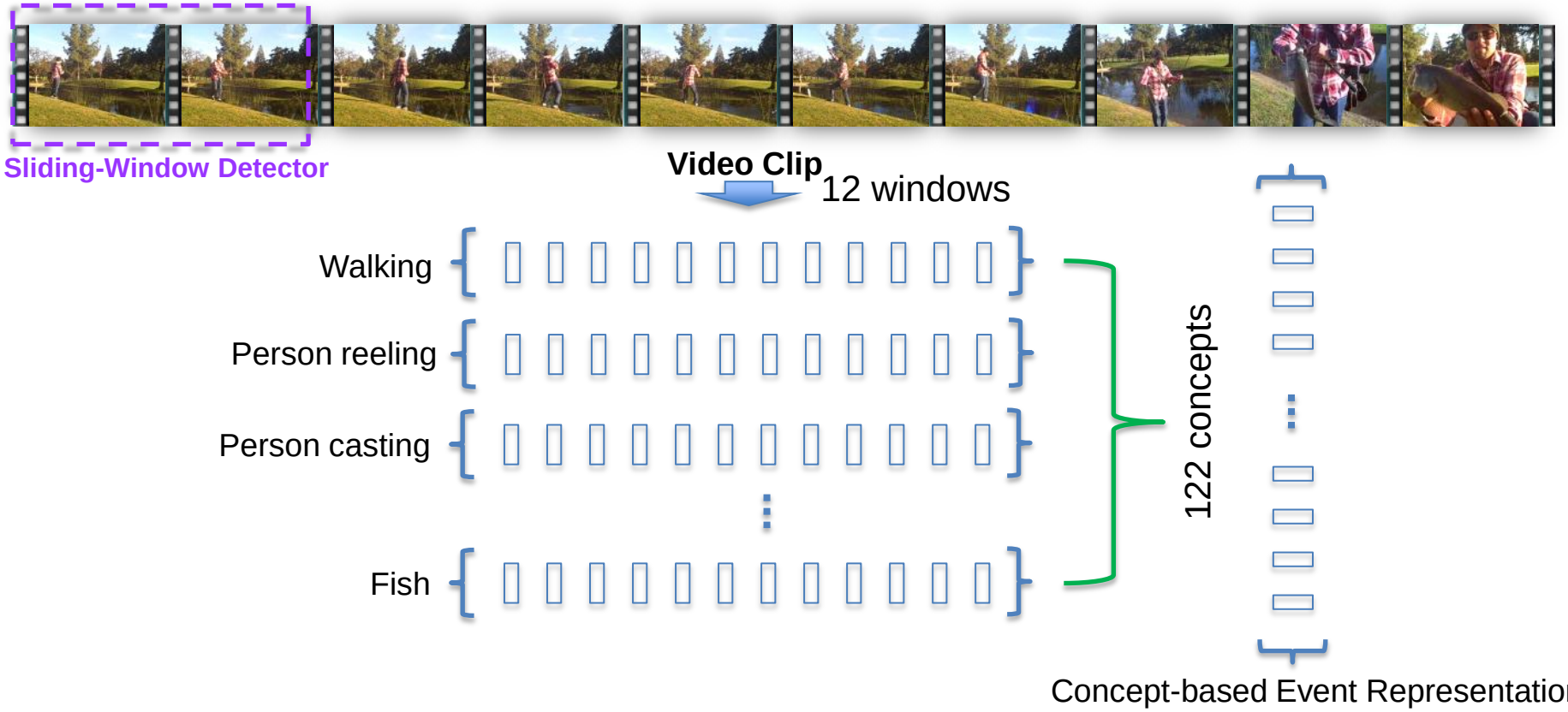
- Concept detectors are mainly trained on static features such as SIFT and color SIFT of key frames

Semantic Concepts

- Part IV: Pseudo Annotation and Object Bank
  - Pseudo Annotation (Umass)
    - 1000 general concepts (still images)
    - Why? Difficult to annotate videos using true concepts
    - Pseudo-annotations are concepts which don't necessarily map to objects
    - Consistent representation
    - Use representation to do video event detection
  - Object Bank (CMU, ICSI)
    - More than 200 object detectors
    - Represent a video as a **Detection Bank**
    - Detection bank consists of a large number of windowed detections with a range of scales
    - Detections are pooled over an image pyramid to provide spatial context
    - Various statistics are computed over the object detections, and their spatial responses

Concept Feature Extraction for Event Detection

- Given a video, a sliding-window based concept detector is applied
- Obtain ***W********N*** detection scores (***W***: number of sliding windows, ***N***: number of concept detectors)



Concept Feature Extraction for Event Detection

- Max_Avg_Std Feature
  - Calculate max, average, and stand deviation of the detection scores
  - Capture the shape of single Gaussian model
- Histogram Vector
  - Compute the frequency (first order) of occurrence of each concepts
  - Capture direct occurrence of concepts in an event
- Co-occurrence Mat
  - Compute the chance of happening between every pair of concepts
  - Captures the overall second order of occurrence
- Max Outer Product
  - Take the max confidence per event and make a outer product
  - Captures the co-occurrence of the most confident concepts

SIN V.S. 81 Action Concepts

- SIN concepts: a large number of concepts defined in a third-party dataset
- 81 action concepts: a small number of concepts defined in MED11 dataset

MD@3%FA	SIN	81-MAX	81-MAS	SIN-81-MAX (EARLY)	SIN-81-MAS (EARLY)	SIN-81-MAX (Late)	SIN-81-MAS (Late)
Birthday_party	0.622	0.715	0.686	0.669	0.593	0.570	**0.558**
Changing_a_vehicle_tire	0.487	0.602	0.646	0.611	0.496	0.434	**0.425**
Flash_mob_gathering	0.200	0.311	0.289	0.267	0.207	0.178	**0.156**
Getting_a_vehicle_unstuck	0.337	0.566	0.554	0.458	0.398	0.349	**0.313**
Grooming_an_animal	0.704	**0.679**	0.716	0.679	**0.568**	0.630	0.580
xMaking_a_sandwich	0.803	**0.781**	0.737	0.788	0.730	0.708	**0.672**
Parade	0.439	0.460	0.417	0.487	0.369	0.305	**0.283**
Parkour	0.382	**0.363**	0.304	0.304	0.265	0.265	**0.245**
Repairing_an_appliance	0.466	**0.420**	0.432	0.455	0.432	**0.341**	0.352
Sewing_project	0.585	0.695	0.646	0.683	0.585	**0.512**	**0.512**
AVERAGE	0.502	0.559	0.543	0.540	0.464	0.429	**0.410**

SIN V.S. 81 Action Concepts

- Confusion Matrix (3%FA)

SIN Concepts

	E06	E07	E08	E09	E10	E11	E12	E13	E14	E15
E06	63	4	3	0	7	11	3	1	1	4
E07	1	57	1	10	7	5	5	0	11	3
E08	3	2	106	0	0	1	34	1	0	2
E09	0	1	1	55	1	2	0	3	2	0
E10	0	6	0	2	23	0	0	1	3	5
E11	3	6	0	0	6	26	2	0	12	10
E12	2	4	47	1	0	0	102	5	0	0
E13	0	2	6	2	2	2	1	62	2	0
E14	0	10	0	2	5	10	0	0	46	11
E15	3	3	0	0	1	7	0	1	4	33
E00	879	855	834	877	865	854	846	881	859	859

81 Action Concepts

	E06	E07	E08	E09	E10	E11	E12	E13	E14	E15
E06	54	2	6	0	3	17	2	2	2	8
E07	0	40	2	5	4	2	7	1	14	2
E08	8	1	96	0	2	1	16	3	0	2
E09	1	4	2	37	0	0	1	0	3	1
E10	0	2	0	0	23	1	0	0	2	0
E11	3	7	2	0	3	35	1	1	22	12
E12	8	5	14	3	3	2	108	5	0	0
E13	0	0	2	4	2	0	0	70	0	0
E14	1	9	0	0	2	5	0	0	49	10
E15	5	0	0	0	2	4	0	0	7	28
E00	866	864	865	883	874	861	865	882	845	860

E06	Birthday_party	E09	Getting_a_vehicle_unstuck	E12	Parade	E15	Sewing_project
E07	Changing_a_vehicle_tire	E10	Grooming_an_animal	E13	Parkour	E00	Background
E08	Flash_mob_gathering	E11	Making_a_sandwich	E14	Repairing_an_appliance		

SIN V.S. 81 Action Concepts

- Confusion Matrix (3%FA)

Late fusion of SIN Concepts and 81 Action Concepts

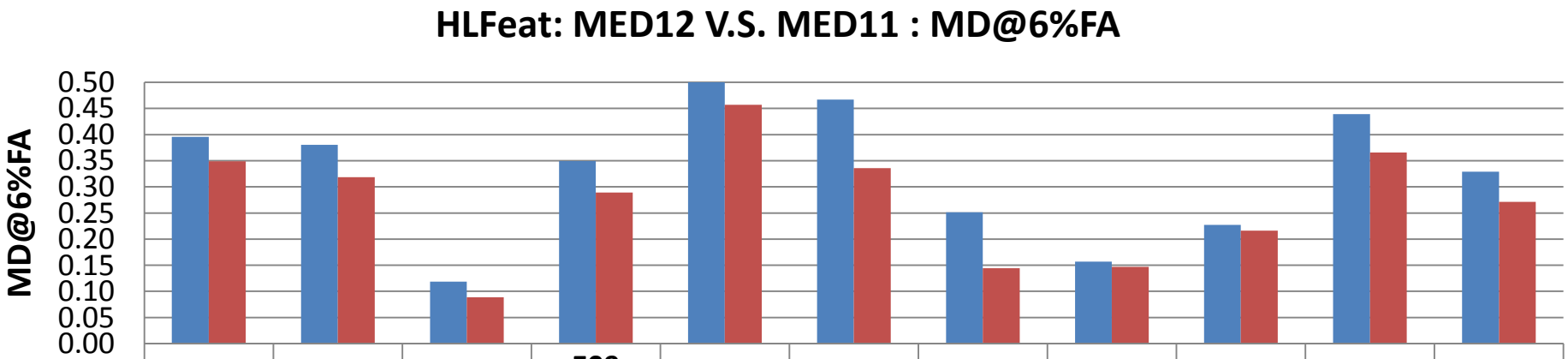
	SIN+MAS Late (3%)	E06	E07	E08	E09	E10	E11	E12	E13	E14	E15
E06	Birthday_party	73	1	5	0	4	20	2	0	1	2
E07	Changing_a_vehicle_tire	0	63	1	6	4	4	6	0	12	3
E08	Flash_mob_gathering	4	2	114	0	0	1	25	0	0	1
E09	Getting_a_vehicle_unstuck	0	6	1	56	0	0	0	1	2	0
E10	Grooming_an_animal	0	1	0	2	32	0	0	1	2	4
E11	Making_a_sandwich	2	10	0	0	4	43	1	0	20	11
E12	Parade	1	8	34	2	1	0	132	2	0	0
E13	Parkour	0	1	4	3	3	0	0	76	1	0
E14	Repairing_an_appliance	0	9	0	1	2	8	0	0	57	12
E15	Sewing_project	3	2	0	0	2	5	0	0	3	38
E00	Background	882	854	848	881	875	856	858	890	854	862

Concept Features Comparison

- Train on EC+DEVT, Test on DEVO
  - **PA**: Pseudo Annotation, **EA**: External Action Concepts = UCF 50 YouTube Actions + BROWN HMDB 51 Actions
  - **MAS, CoMat, Hist, MaxCoMat** are computed from MED11 81 Action Concepts
  - **SIN**: MED11 346 concepts
- Conclusions:
  - HIST and EA are not helpful.
  - PA, SIN and MAS+CoMat+MaxCoMat on 81 actions are complementary concept features.

MD@6%FA	PA (SVM-Rank)	Ex-Actions (EA) (RBF)	SIN (RBF)	MAS (RBF)	CoMat	HIST	maxCoMat (RBF)	CoMat + MaxCoMat	CoMat +MaxCoMat+ HIST	CoMat +MaxCoMat+ MAS	81Action+SIN	81Action+SIN+ EA
Birthday_party	0.49	0.75	0.53	0.58	0.52	0.53	0.47	0.42	0.42	0.43	0.34	0.35
Changing_a_vehicle_tire	0.32	0.81	0.48	0.51	0.38	0.48	0.49	0.35	0.40	0.35	0.35	0.38
Flash_mob_gathering	0.08	0.30	0.18	0.19	0.21	0.24	0.19	0.15	0.13	0.13	0.08	0.08
Getting_a_vehicle_unstuck	0.28	0.59	0.42	0.30	0.31	0.39	0.42	0.34	0.31	0.30	0.31	0.31
Grooming_an_animal	0.54	0.83	0.74	0.65	0.56	0.56	0.62	0.48	0.53	0.49	0.44	0.42
Making_a_sandwich	0.46	0.61	0.66	0.71	0.50	0.55	0.55	0.48	0.47	0.48	0.43	0.42
Parade	0.37	0.61	0.36	0.42	0.38	0.41	0.28	0.22	0.24	0.21	0.17	0.19
Parkour	0.33	0.42	0.50	0.35	0.21	0.25	0.21	0.17	0.17	0.19	0.14	0.16
Repairing_an_appliance	0.28	0.68	0.39	0.33	0.35	0.41	0.42	0.31	0.30	0.28	0.26	0.24
Sewing_project	0.43	0.78	0.66	0.60	0.52	0.54	0.50	0.48	0.44	0.44	0.44	0.44
AVERAGE	0.36	0.64	0.49	0.46	0.39	0.43	0.41	0.34	0.34	0.33	0.30	0.30

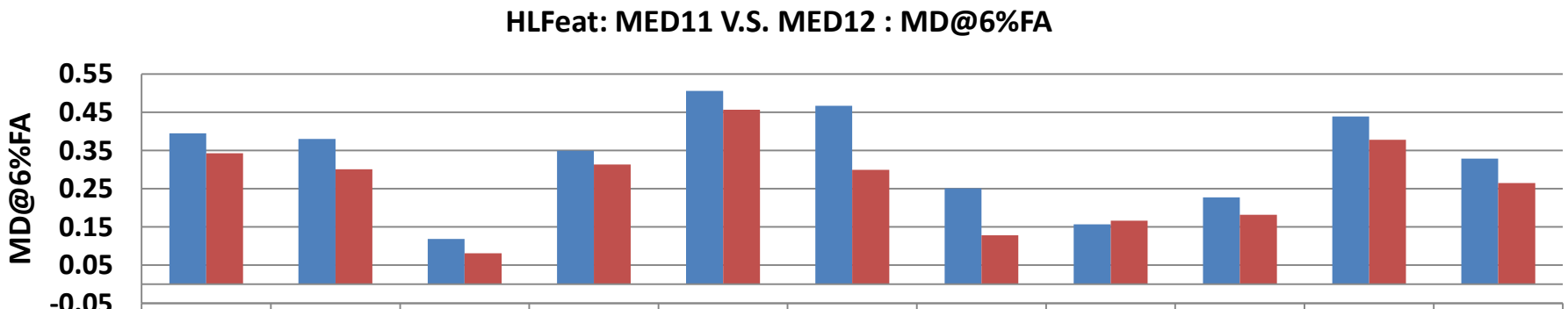
MED12 V.S. MED11: High-Level Concept Features *without OCR / ASR*



	E06 Birthday_party	E07 Changing_a_vehicle_tire	E08 Flash_mob_gathering	E09 Getting_a_vehicle_unstuck	E10 Grooming_an_animal	E11 Making_a_sandwich	E12 Parade	E13 Parkour	E14 Repairing_an_appliance	E15 Sewing_project	Average
MED11	0.40	0.38	0.12	0.35	0.51	0.47	0.25	0.16	0.23	0.44	0.33
MED12	0.35	0.32	0.09	0.29	0.46	0.34	0.14	0.15	0.22	0.37	0.27

- MED11 Semantic Concepts
  - 81 Action Concepts + 20 Audio Concepts + 17 Scene/Object Concepts
- MED12 Semantic Concepts
  - 185 Action Concepts + 346 SIN Concepts + 1000 PA Concepts + *167 Object Bank*

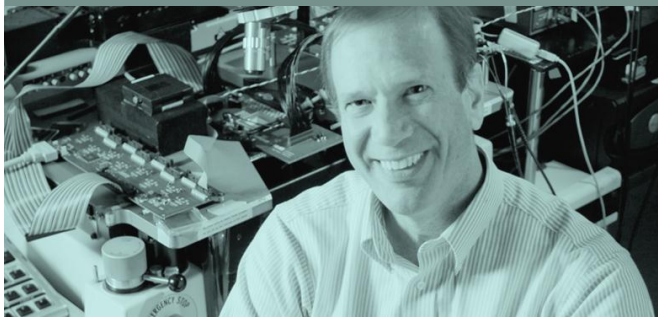
MED12 V.S. MED11: High-Level Concept Features *with OCR / ASR*



	E06 Birthday_party	E07 Changing_a_vehicle_tire	E08 Flash_mob_gathering	E09 Getting_a_vehicle_unstuck	E10 Grooming_an_animal	E11 Making_a_sandwich	E12 Parade	E13 Parkour	E14 Repairing_an_appliance	E15 Sewing_project	Average
MED11	0.40	0.38	0.12	0.35	0.51	0.47	0.25	0.16	0.23	0.44	0.33
MED12	0.34	0.30	0.08	0.31	0.46	0.30	0.13	0.17	0.18	0.38	0.26

- MED11 Semantic Concepts
  - 81 Action Concepts + 20 Audio Concepts + 17 Scene/Object Concepts
- MED12 Semantic Concepts
  - 185 Action Concepts + 346 SIN Concepts + 1000 PA Concepts + *167 Object Bank*

SRI International
SARNOFF



MED’12 Evaluation Results

Summary on MED12 Evaluation

- 20 Pre-specified Events

EK-FULL Runs			
	NDC	**FA**	**MD**
c-LLFeature	0.640	0.030	0.268
c-HLFeatAsrOcr	0.745	0.034	0.323
p-LLFeatHLFeatAsrOcr	0.642	0.031	0.261
EK10Ex Run			
c-EK10xLLFeatHLFeatAsr	0.891	0.040	0.395

  - LLFeat-Run: low-level features based run
  - HLFeat-ASR/OCR_Run: concept features plus ASR/OCR based run
  - LLFeat-HLFeat-ASR/OCR_Run (our primary run) : the combination of low-level features, high-level features and ASR/OCR.

– 5 AdHoc Events (Actual Decision)

EK-FULL Runs			
	NDC	**FA**	**MD**
p-LLFeatHLFeatAsrOcr	0.641	0.027	0.299
EK10Ex Run			
c-EK10xLLFeatHLFeatOcrAsr	1.010	0.043	0.470

Acknowledgement

This work has been supported by the **Intelligence Advanced Research Projects Activity (IARPA)** via Department of Interior National Business Center contract number D11-PC20066.

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not with-standing any copyright annotation thereon. 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.