Raytheon
BBN Technologies

Video Indexing and Search with Event Recounting (VISER)

TRECVID 2013

Presented by

Shuang Wu

Scientist, Raytheon BBN Technologies

BBN VISER at TRECVID 2013

- Participated in both MED and MER tasks
- Made submissions for all event/training/system conditions
- Continue to build and improve upon core system work from previous years in TRECVID
  - Multi-modal feature extraction
  - Max-margin classification and multi-stage fusion
- One major area of focus in 2013: **semantics**

Semantics for MED and MER

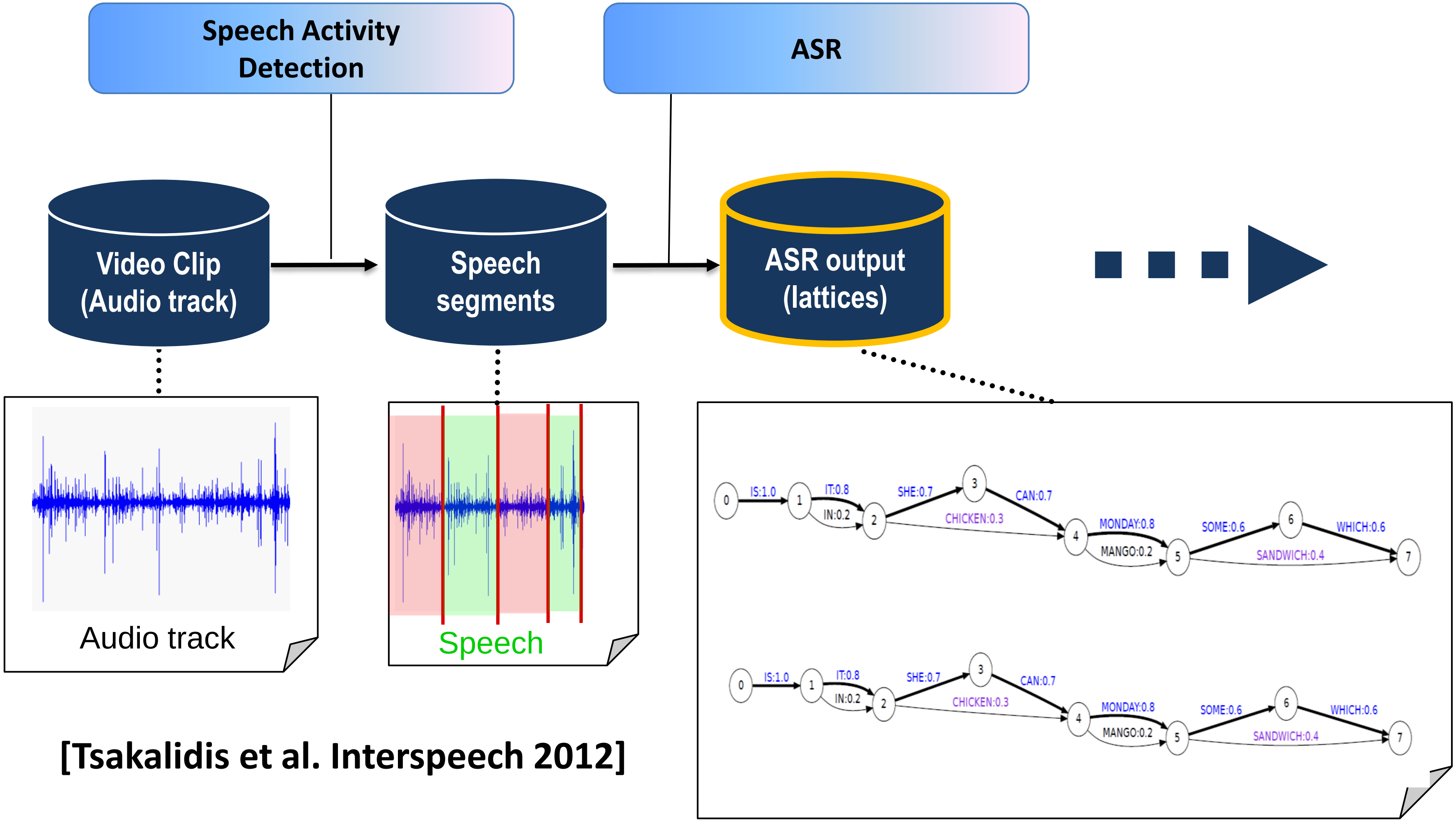
- Increasing necessity in TRECVID for semantic understanding of video
  - **MER**: semantic explanation of event detection
  - **MED 0Ex**: video event detection from text query only
- **Key building block for both problems**: reliable semantic extraction from video

Overview

- **Language extraction**:
  - Speech and video text
- **Audio-visual concepts**:
  - Off-the-shelf detectors
  - In-domain detectors
- **Experiments**:
  - 0Ex MED
  - Semantics in 10Ex/100Ex MED
- **TRECVID 13 results:**
  - MED
  - MER

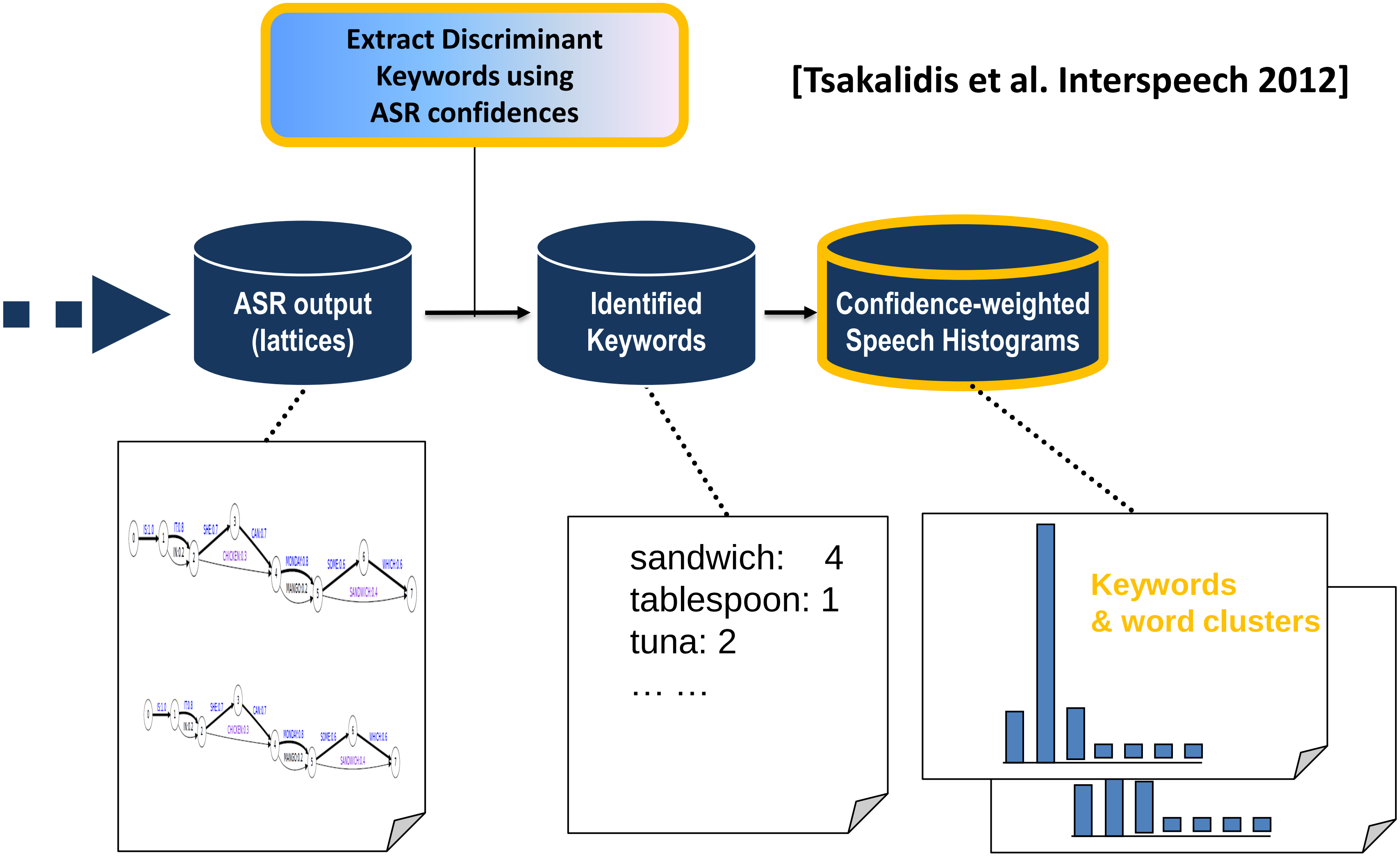
Language Extraction

Speech



[Tsakalidis et al. Interspeech 2012]

Speech (cont'd)



Extract Discriminant Keywords using ASR confidences

[Tsakalidis et al. Interspeech 2012]

ASR output (lattices)

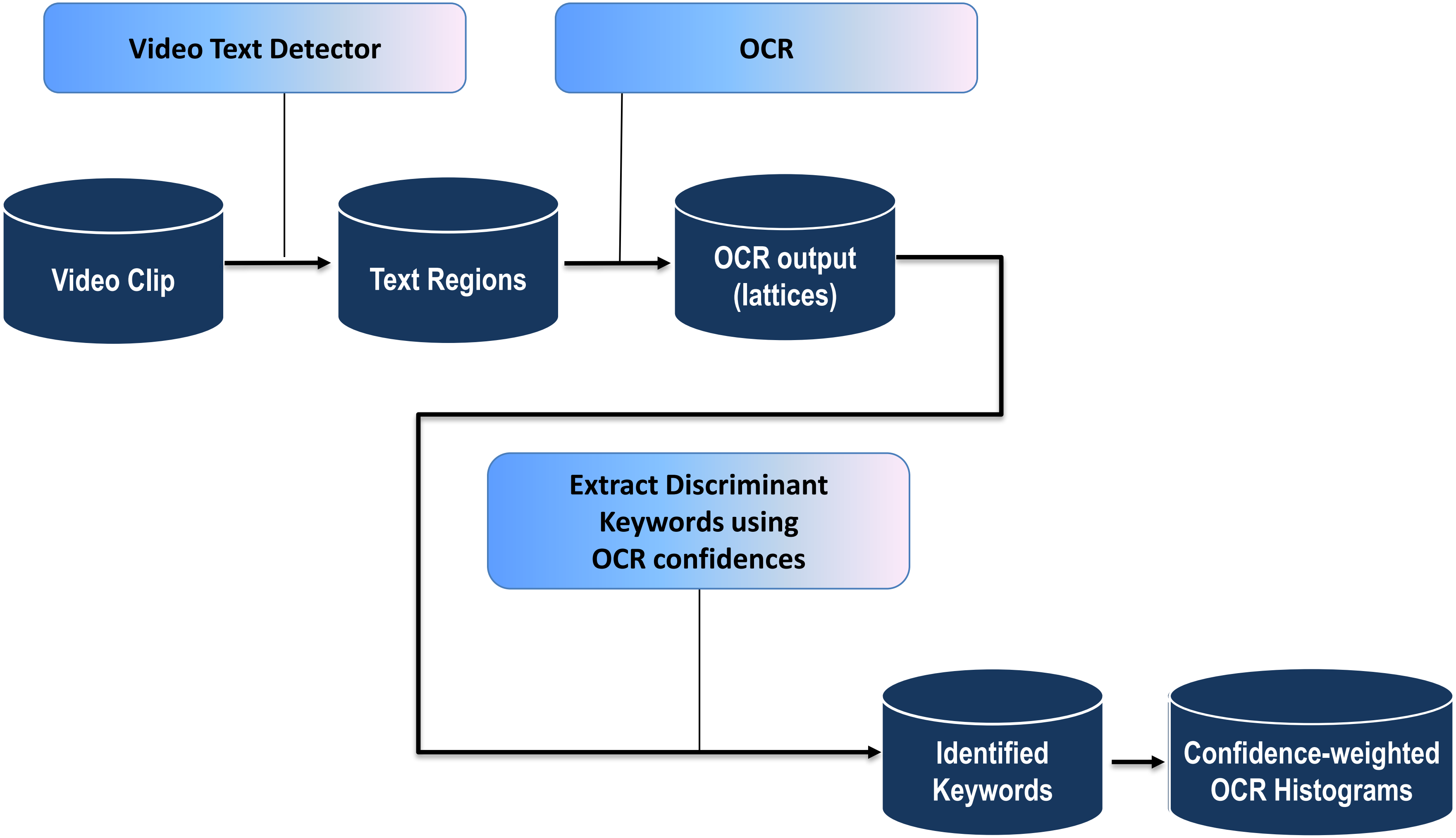
Identified Keywords

Confidence-weighted Speech Histograms

sandwich: 4
tablespoon: 1
tuna: 2
… …

Keywords & word clusters

Video Text


Video Text Detector
OCR
Video Clip
Text Regions
OCR output (lattices)
Extract Discriminant Keywords using OCR confidences
Identified Keywords
Confidence-weighted OCR Histograms

Language Content Frequency

- Keyword detections are usually precise
- Only a third of the data has relevant speech, and even less has video text
- Relevant speech and text content in web video is too sparse…

Audio-visual Concepts

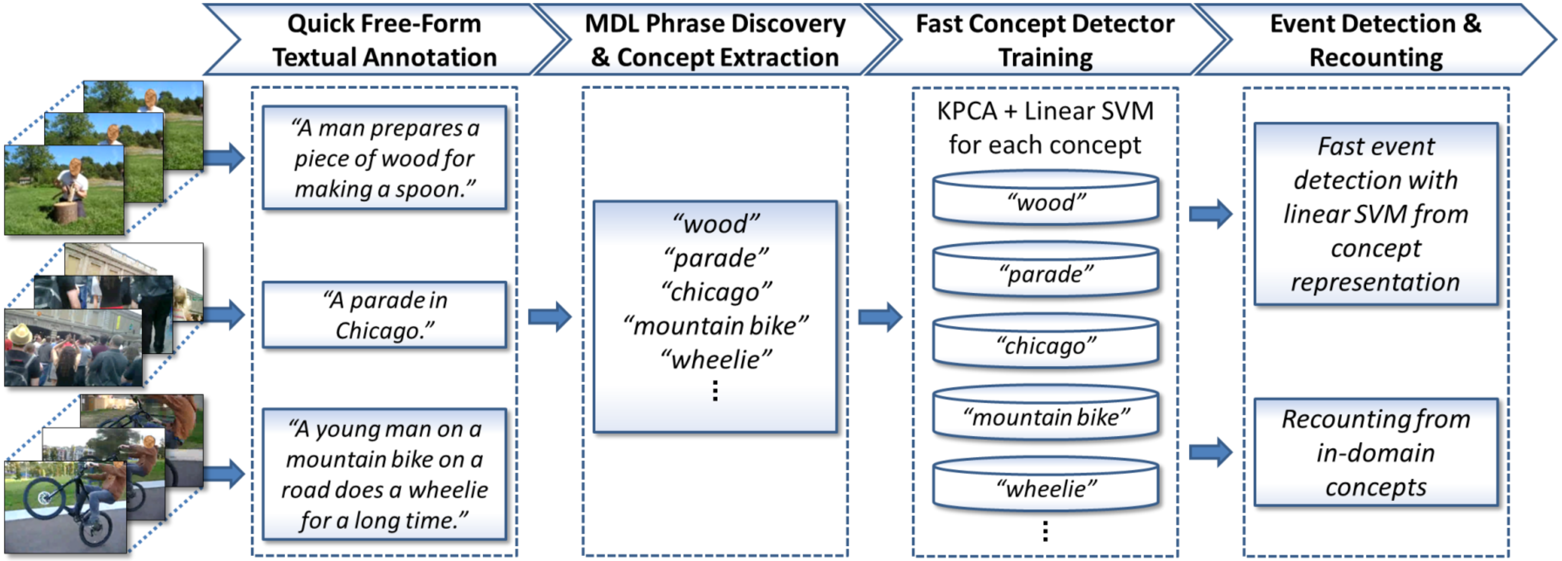
Off-the-shelf Concept Detectors

- Evaluated several off-the-shelf concept detector collections
  - Classemes [Torresani *et al.*, ECCV ‘10]
  - ObjectBank [Li *et al.*, NIPS ‘10]
  - Sun Scene [Patterson *et al.*, CVPR ‘12]
- Domain mismatch
  - Image vs. video; professional vs. user-contributed quality
  - Pre-defined concept ontology vs. MED/MER concepts
- Could address issues with adaptation, but…

In-domain Concept Discovery

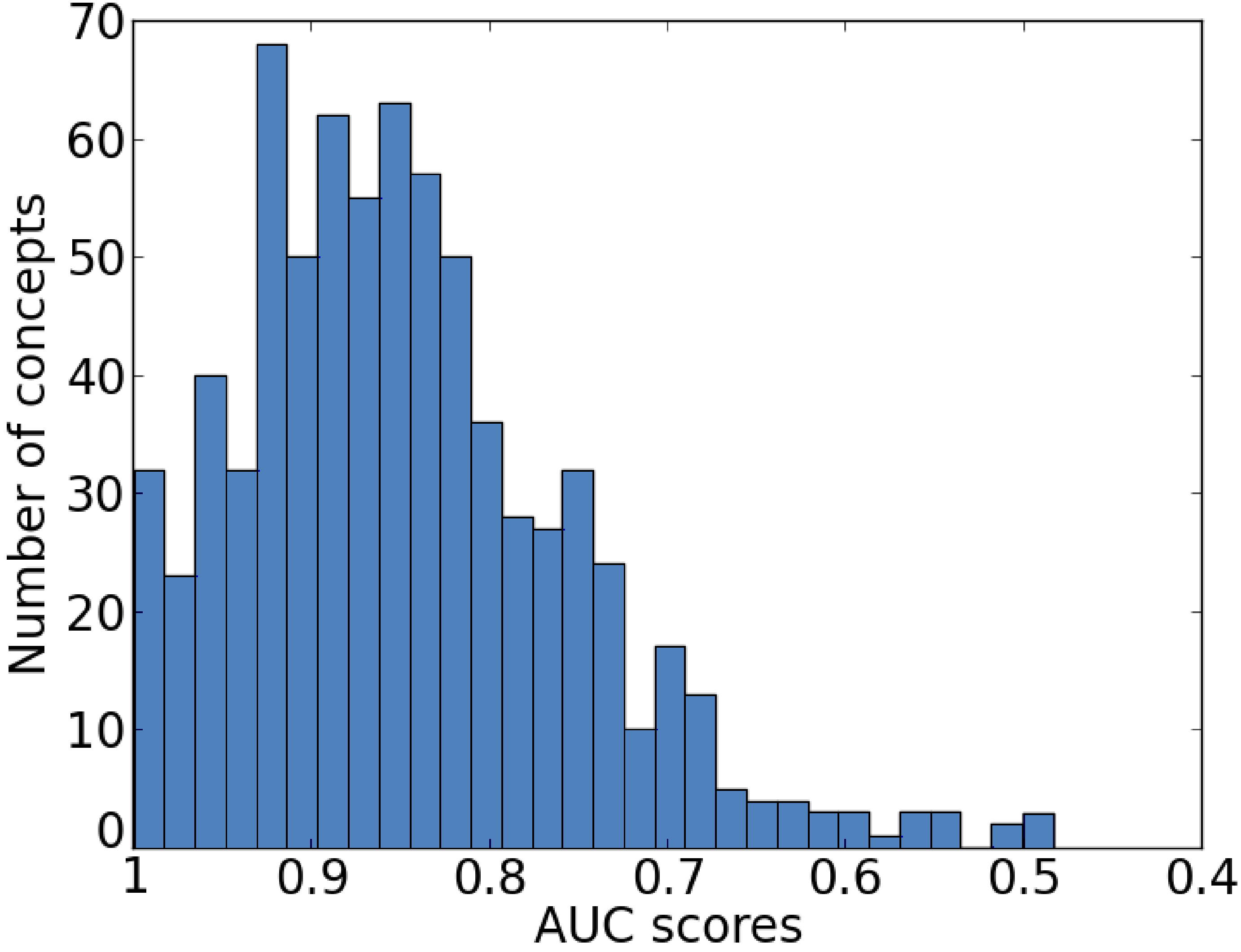
- Start with in-domain data: **MED research collection**
- Minimized domain mismatch but no concept annotation
- Available short text summaries in judgement files
- Discover concept labels from natural language snippets
  - Efficient to collect: **28x faster** than annotating fixed concept ontology
  - No predefined constraints on concept vocabulary

Weakly Supervised Concepts (WSC)


Quick Free-Form Textual Annotation
MDL Phrase Discovery & Concept Extraction
Fast Concept Detector Training
Event Detection & Recounting
"A man prepares a piece of wood for making a spoon."
"A parade in Chicago."
"A young man on a mountain bike on a road does a wheelie for a long time."
"wood"
"parade"
"chicago"
"mountain bike"
"wheelie"
⋮
KPCA + Linear SVM for each concept
"wood"
"parade"
"chicago"
"mountain bike"
"wheelie"
⋮
Fast event detection with linear SVM from concept representation
Recounting from in-domain concepts

- Natural language pre-processing and phrase discovery
- Leverage existing MED infrastructure and extracted concept labels to train concept detectors
- Concept selection via cross-validation

Concept Performance Distribution



Examples of Top Concepts Detected

Event: birthday party (E006)
WSC recounting: piñata, people celebrate, gift, surprise party.
Classemes: chemical weapon, collection, display setting, backpacker

Event: changing a vehicle tire (E007)
WSC recounting: tire, change, replace, technique.
Classemes: chemical weapon, physical creation event, dangerous activity, movement translation process

Event: flash mob gathering (E008)
WSC recounting: dance flash mob, shopping, hong kong
Classemes: chemical weapon, collection display setting, small group, windsurfer

Event: getting a vehicle unstuck (E009)
WSC recounting: rocky, jeep, trail, car.
Classemes: collection display setting, anti-armor mine, mine, fighting hole

Event: grooming an animal (E010)
WSC recounting: dog, carve, bathe, pig.
Classemes: chemical weapon, collection display setting, single doer action, diplomat

WSC Concept Flexibility

- We can train WSC concepts using any of the existing features/modalities already present in the MED infrastructure
  - Can learn visual and audio features from the same discovered labels, as well as multi-modal detectors
- Initial exploration of web data download allowed in TRECVID 13:
  - Retrieved image data via Google Image Search and Youtube Search API (thumbnail only)
  - Train visual WSC features

Concept Distance (CD) Detectors

- SVM training is unreliable for concepts with few examples
  - e.g. 10Ex MED
- Alternative fast and simple concept detection approach
- Using discovered concepts *C*, V$_c$ videos for each concept *c* in *C*
- Compute Concept Distance (CD) model $\mathbf{y}_c$ (vector) by simple aggregation of feature vectors in V$_c$
- Concept detection score is computed as distance of test video to $\mathbf{y}_c$

Experiments

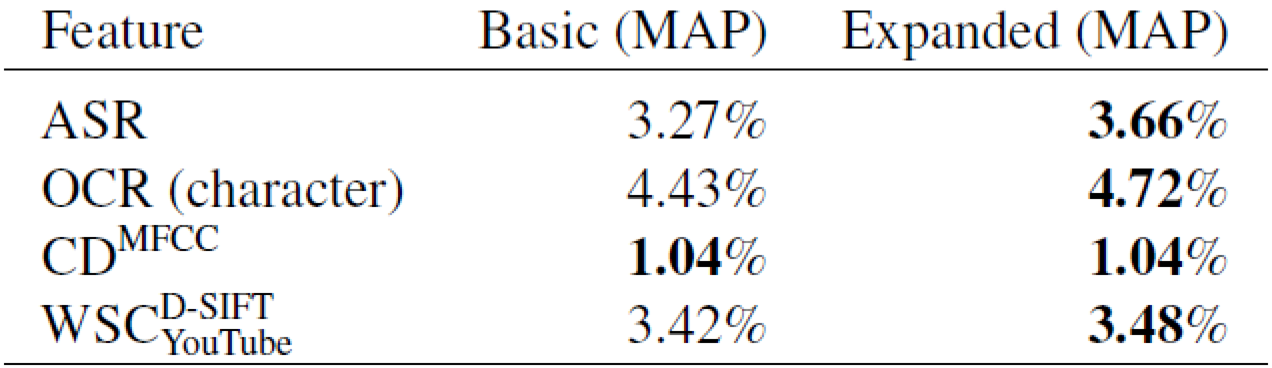
0Ex MED

- Convert test video to semantic concepts
  - Off-the-shelf detectors
  - WSC
  - CD
  - ASR/OCR keyword spotting
- Convert event-kit description (query) to concept list
- Improve query-video concept matching with text expansion
- Compute query-video similarity scores

Experimental Setup

- Tested on 20 events in MEDTest collection (~27k videos)
- WSC/CD learning on Research set (~10k videos)
- Use Event Kit descriptions (~250 words length) as queries

Text Expansion

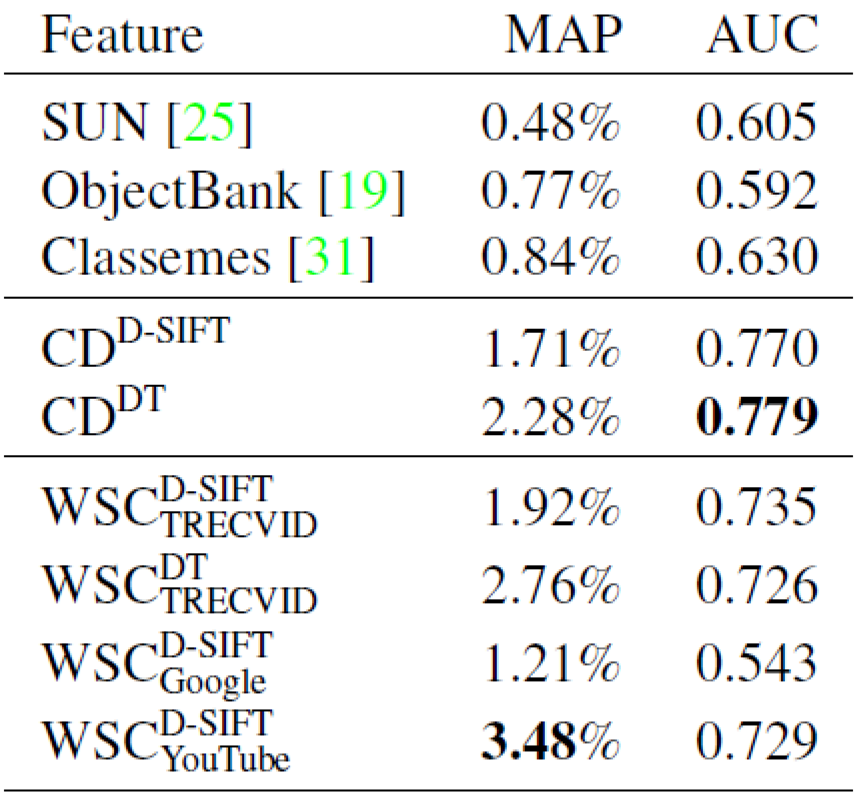


Feature	Basic (MAP)	Expanded (MAP)
ASR	3.27%	**3.66%**
OCR (character)	4.43%	**4.72%**
CD^{MFCC}	**1.04%**	**1.04%**
$\text{WSC}^{\text{D-SIFT}}_{\text{YouTube}}$	3.42%	**3.48%**

- Small but consistent gains with text expansion

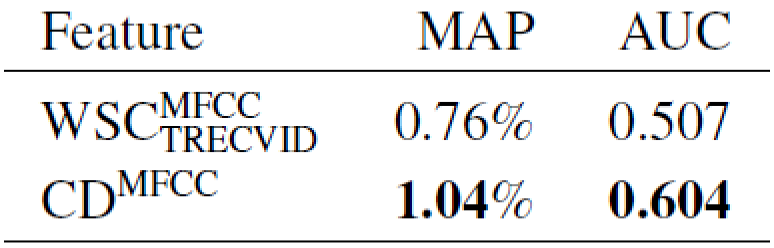
Visual Concepts

- Off-the-shelf detectors have poor performance
- CD features strong despite simplicity
- $\text{WSC}_{\text{YouTube}}$ has best performance



Feature	MAP	AUC
SUN [25]	0.48%	0.605
ObjectBank [19]	0.77%	0.592
Classemes [31]	0.84%	0.630
$\text{CD}^{\text{D-SIFT}}$	1.71%	0.770
CD^{DT}	2.28%	**0.779**
$\text{WSC}^{\text{D-SIFT}}_{\text{TRECVID}}$	1.92%	0.735
$\text{WSC}^{\text{DT}}_{\text{TRECVID}}$	2.76%	0.726
$\text{WSC}^{\text{D-SIFT}}_{\text{Google}}$	1.21%	0.543
$\text{WSC}^{\text{D-SIFT}}_{\text{YouTube}}$	**3.48%**	0.729

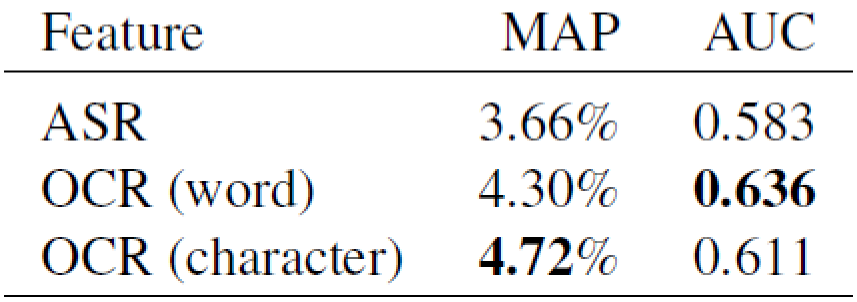
Audio Concepts



Feature	MAP	AUC
WSC$^{\text{MFCC}}_{\text{TRECVID}}$	0.76%	0.507
CD$^{\text{MFCC}}$	**1.04%**	**0.604**

- Suffers from sparse content and unrelated audio content

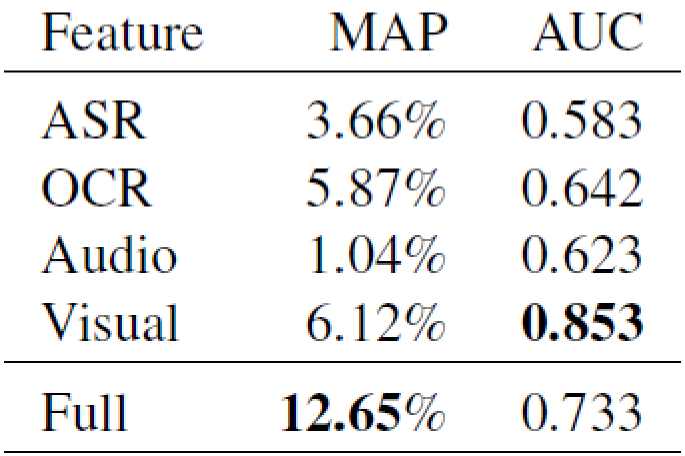
Language Keywords



Feature	MAP	AUC
ASR	3.66%	0.583
OCR (word)	4.30%	**0.636**
OCR (character)	**4.72%**	0.611

- All systems have higher MAP than audio-visual concepts
- Lower AUC: sparse language content

Fusion



Feature	MAP	AUC
ASR	3.66%	0.583
OCR	5.87%	0.642
Audio	1.04%	0.623
Visual	6.12%	**0.853**
Full	**12.65%**	0.733

- WSC/CD concepts are complementary
  - Off-the-shelf detectors discarded due to negative performance impact
- Fusion across modalities more than doubles individual performance

10Ex/100Ex MED

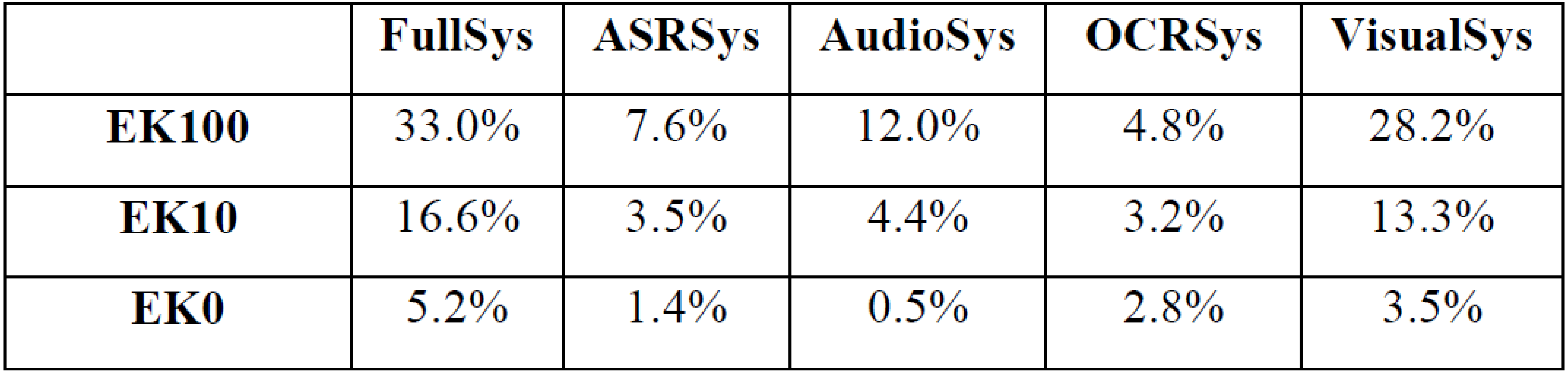
- Measure usefulness of visual semantic information on top of low-level visual information
- Treat concept scores as feature vector in standard 10Ex/100Ex MED framework
- Semantics particularly helpful in 10Ex

Training/Features	MAP	R0
10Ex/LLFeat	0.1459	0.1885
10Ex/LLFeat+WSC	**0.1785**	**0.2190**
100Ex/LLFeat	0.3810	0.4771
100Ex/LLFeat+WSC	**0.3852**	**0.4830**

TRECVID 13 Results

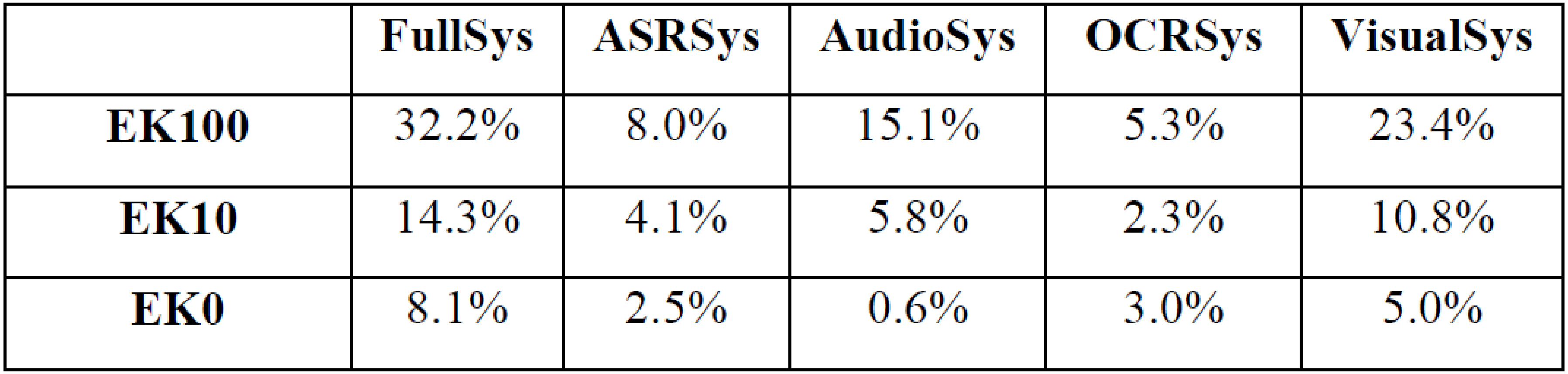
MED

Prespecified



	FullSys	ASRSys	AudioSys	OCRSys	VisualSys
EK100	33.0%	7.6%	12.0%	4.8%	28.2%
EK10	16.6%	3.5%	4.4%	3.2%	13.3%
EK0	5.2%	1.4%	0.5%	2.8%	3.5%

Ad Hoc



	FullSys	ASRSys	AudioSys	OCRSys	VisualSys
EK100	32.2%	8.0%	15.1%	5.3%	23.4%
EK10	14.3%	4.1%	5.8%	2.3%	10.8%
EK0	8.1%	2.5%	0.6%	3.0%	5.0%

MED

- Consistent prespecified and ad hoc performance
  - Our in-domain concepts are event-independent and generalize well to different event queries
- Strong overall performance in all system conditions

MER

- **Philosophy:** fine balance between presenting enough information, and keeping review times short enough
- Filter concept detections with low confidence or relevance to detected event
- Aggregate semantic information by modality, taking into account temporal overlap to merge evidences
- Generate short list of itemized evidences sorted by confidence and relevance

MER

- Overall accuracy: 64.96%
- Percent recounting review time: 50.59%
- Observation text precision: 1.78

Summary

- Reliable semantic extraction from video is key for many MED/MER tasks
- Leveraging in-domain data produces strong results even from simple methods
- Multi-modal combination of semantic information is especially important
- Semantics can contribute even to traditional 10Ex/100Ex MED

Acknowledgement

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding 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.

Thank You!