Raytheon
BBN Technologies

# BBN VISER Zero-Shot Multimedia Event Detection

Manasvi Tickoo, Florian Luisier, and Walt Andrews
*Speech, Language, and Multimedia Business Unit*
*Raytheon BBN Technologies, Cambridge MA*

## Key Contributions

- Automated framework for zero-shot learning; i.e., high-level multimedia event detection with no positive training exemplars
- Automated discovery and learning of concept detectors by leveraging free-form text descriptions of videos or images
- Automatically generated semantic query based on a free-form description of an event (or just the event name) and INDRI Document Retrieval System. Score words based on their TF scores from the top 'N' documents
- Multiple fusion steps to achieve optimal performance
- Strong results on TRECVID MED14Test dataset

## Multimodal Concept Detectors

- In-domain Weakly Supervised Concepts (WSC):
  - Extract low-level features:
    D-SIFT, O-SIFT, Dense Trajectories, and MFCC features with Fisher vector encoding
  - Collect free-form text description of videos from research set and YouTube
  - Use NLP techniques to extract weak video-level concept labels
  - Train SVM-based linear classifiers from weak labels, using cross-validation to tune parameters and prune weak concepts
  - Use vector of detection probabilities as WSC vector (1,800)
- Deep Convolutional Neural Network (DCNN) features trained on the ImageNet dataset: 1,000 object detectors
- Automatic Speech Recognition (ASR)
- Optical Character Recognition (OCR)

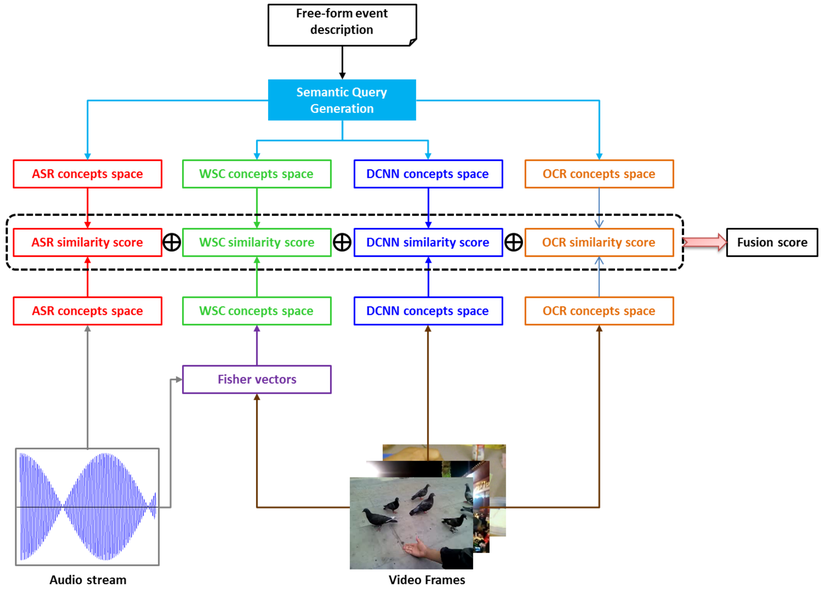
## Similarity Computation

- Similarity of projected query vector $\mathbf{f}_Q$ and video vector $\mathbf{f}_v$ computed as:

$$S_Q(v) = \mathbf{f}_Q^T \mathbf{f}_v$$

- Max-min normalization before late average fusion of multiple modalities

## System Overview




## Event Query Projection

- Expand query ($Q$) and project into vocabulary ($V$) :

**For each** word $v$ in $V$, do
**If** $v \in Q$, then $\text{score}(v) = 1$ **End**
**If** $v \notin Q$, then
**For each** $w \in Q$, expand $w$ into $W = \{w_1, \ldots, w_k\}$ using Gigaword similarity matrix. Then,
**For each** $w_k$ in $W$, **do**
$\text{score}(v)$ += $\text{sim}(v, w_k)$
**End**
**End**
**End**
**End**

## TRECVID MED14 Experiments

- **Concept training set:**
  TRECVID Research Set + YouTube (about 20,000 videos)
- **Event training set:**
  TRECVID Event Background Set (about 5,000 videos)
- **Test set:**
  TRECVID MED14Test Set (about 25,000 videos) E021-E040

| Feature | ASR | Audiovisual WSC | DCNN | OCR | Fusion |
|---|---|---|---|---|---|
| MAP | 0.027 | 0.044 | 0.018 | 0.034 | **0.090** |
| MR0 | 0.034 | 0.155 | 0.014 | 0.056 | **0.250** |
| AUC | 0.617 | 0.890 | 0.764 | 0.609 | **0.923** |

**Top Retrievals for the query "Attempting a bike trick"**

## Conclusions

- Video mapped to intermediate semantic attribute space through multiple multi-modal features
- Proposed in-domain concept learning more robust than off-the-shelf detectors
- Disjoint vocabularies between query and various modalities aligned through query expansion
- Final system significantly better than any individual feature or modality thanks to simple fusion techniques

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