Multimodal Sparse Coding for Event Detection

Youngjune Gwon, William M. Campbell, Kevin Brady, Douglas Sturim – MIT Lincoln Laboratory   Miriam Cha, H.T. Kung – Harvard University

Overview

Multimedia Event Detection (MED)

- Aims to identify complex activities consisting of various human actions and objects at different places and times

Motivation

- Given the accelerated growth of multimedia data on the Web (e.g., Facebook, Youtube, and Vimeo), the ability to identify complex activities in the presence of diverse modalities is becoming increasingly important

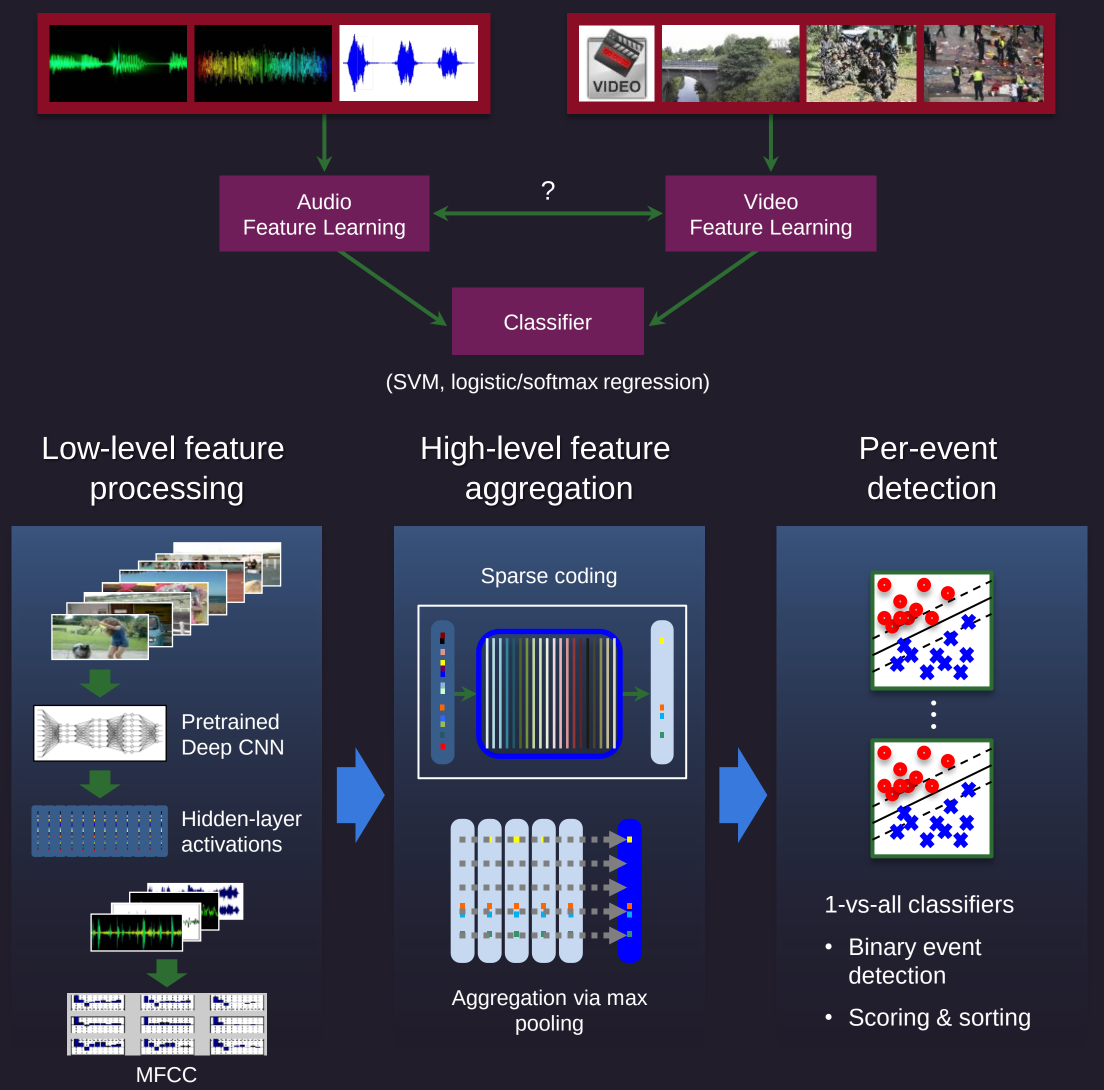
Approach

- Sparse coding-based framework that can model semantic correlation between modalities
  - Sparse coding has been used widely in machine learning applications (e.g., classification, denoising, and recognition) for multimedia data
- Our framework can learn multimodal features by forcing shared sparse code representation between multiple modalities

Result

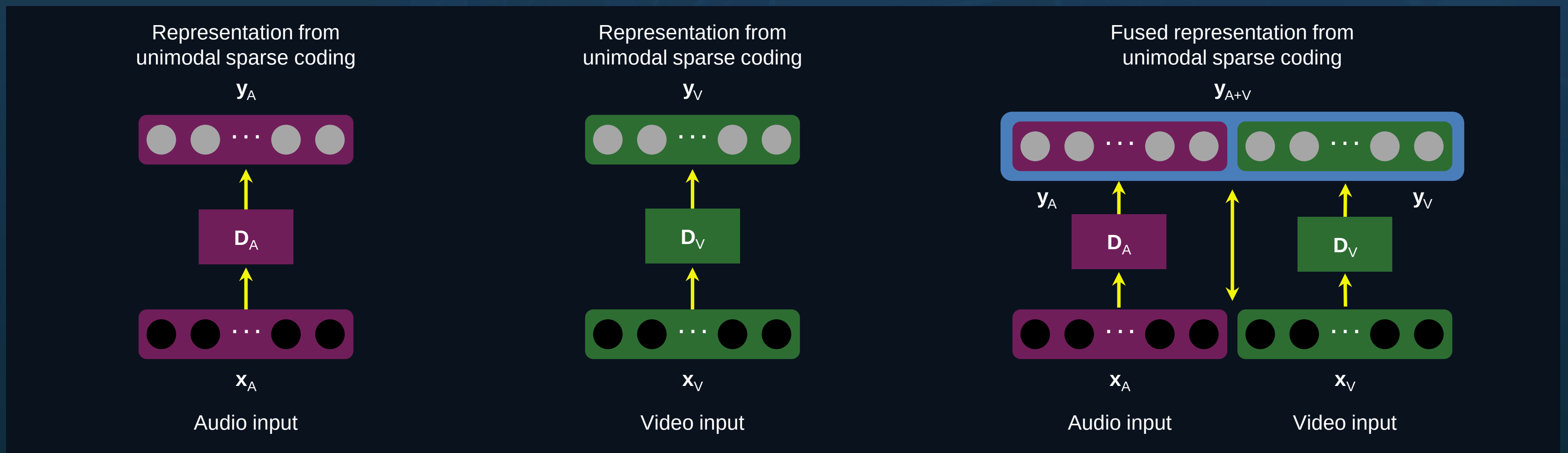
- We present joint feature learning methods that can go beyond simple concatenation of unimodal features
- Our models are validated on TRECIVD dataset, demonstrating competitive audio-video based multimedia event detection

Multimodal Feature Learning



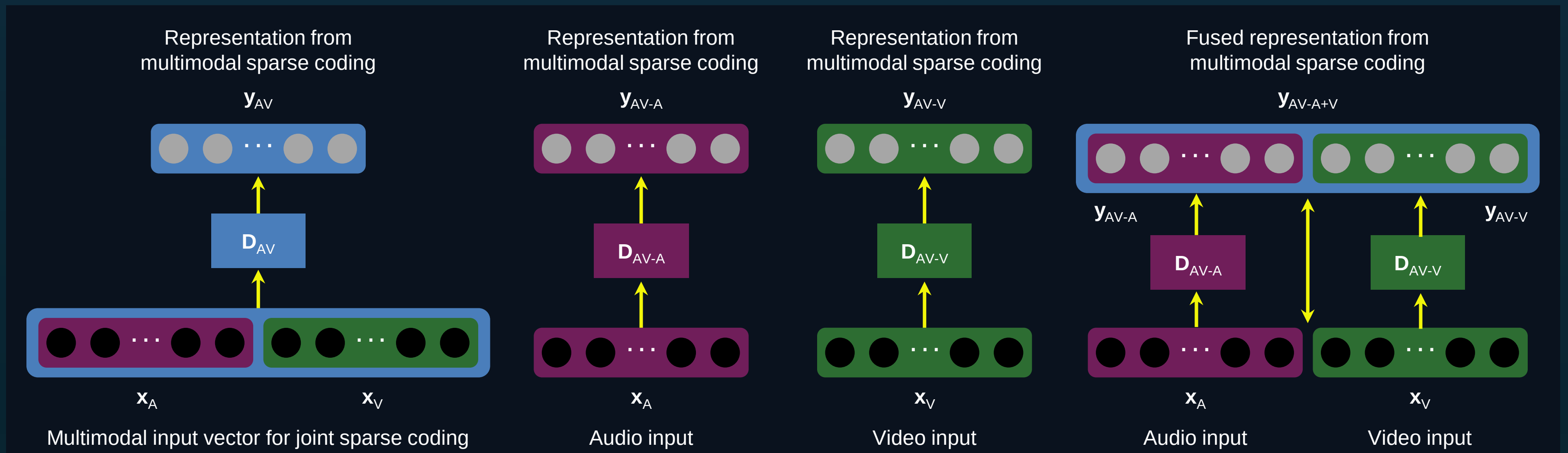
Approach

Approach 1: Unimodal Feature Learning



(a) Audio only (b) Video only (c) Union of unimodal features

Approach 2: Multimodal Joint Feature Learning



(a) Joint sparse coding (b) Cross-modal audio (c) Cross-modal video (d) Union of cross-modal features

Comparative Advantage to Approach 1		
✓ Joint feature learning ⇒ exploit correlation between different modalities	✓ Cross-modality search & retrieval	✓ Novel usages (e.g., McGurk effect, lip sync, talking heads)

Experiments

TRECVID

- Workshop series by NIST since 2001 to promote audio-video analysis and exploitation
- Tasks include MED, semantic indexing, surveillance, instance search

Pipeline

→ Multimedia files → Audio-video data processing → Sparse coding → Pooling → Classifier →

Evaluation

NIST TRECVID MED 2014
- 20 event classes (E021–E040)
- 10Ex and 100Ex data scenarios

Experiments
- Cross-validation on 10Ex
- Train on 10Ex and test with 100Ex

Metrics
- Average 1-vs-all classification accuracy
- Mean average precision (mAP)

Results

Enhancement Resulting from Multimodal Learning

	Unimodal			Multimodal			
	A-only	V-only	Union	A	V	Joint	Union
Mean accuracy (c.v. 10Ex)	69%	86%	**89%**	75%	87%	90%	**91%**
mAP (c.v. 10Ex)	20.0%	28.1%	**34.8%**	27.4%	33.1%	35.3%	**37.9%**
Mean accuracy (10Ex/100Ex)	56%	64%	**71%**	58%	67%	71%	**74%**
mAP (10Ex/100Ex)	17.3%	28.9%	**30.5%**	23.6%	28.0%	28.4%	**33.2%**

- Union of unimodal audio and video feature vectors perform better than using only unimodal features
- Joint sparse coding is able to learn multimodal features that go beyond simply concatenating the two unimodal features
- When the cross-modal features by audio and video are concatenated, they outperform the other feature combinations

Comparison with GMM and RBM

Feature learning schemes	Mean accuracy	mAP
Union of unimodal GMM features	66%	23.5%
Multimodal joint GMM feature	68%	25.2%
Union of unimodal RBM features	70%	30.1%
Multimodal joint RBM feature	72%	31.3%

- Our results show that sparse coding is better than GMM by 5–6% in accuracy and 7–8% in mAP
- Performance of RBM is better than GMM but worse than sparse coding

Summary

Our sparse coding-based approach

- Capable of jointly training mid-level audio (e.g., MFCC) and video (e.g., hidden activations of CNN) features
- Can scale to form file-level feature vector for MED task
- Outperforms GMM and RBM of similar configuration

Future Work

- Use static frames with optical flow for video processing
- Investigate joint feature learning scheme for RBM

This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY