TOWARD VISUAL VOICE ACTIVITY DETECTION FOR UNCONSTRAINED VIDEOS

Rahul Sharma, Krishna Somandepalli and Shrikanth Narayanan
University of Southern California, USA

Problem: Visual VAD

Voice activity detection (VAD) from just the visual modality without the need for face detection and tracking.

Idea

Classify voice activity from video image frames ⇒ Such a model can learn to attend to speaking faces.

Approach

Supervised cross modal learning to map video image frames to audio VAD labels.

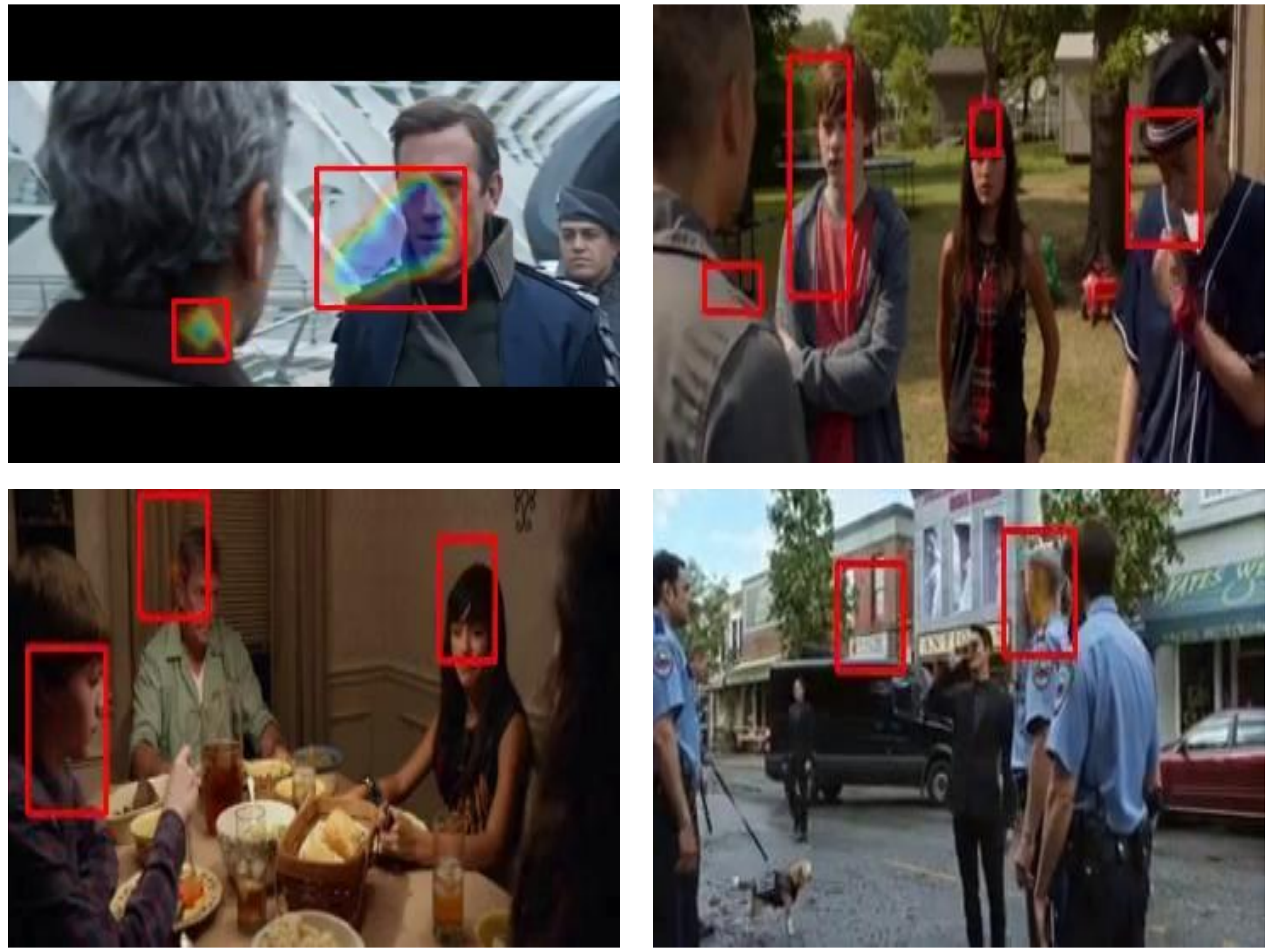
- ❑ We obtain coarse VAD labels from movies' subtitles.
- ❑ We propose **Hierarchical Context Aware (HiCA)** deep architecture that can capture short-term spatial-temporal context and long-term temporal context.
- ❑ The use of 3D CNNs makes HiCA highly interpretable.
- ❑ We show that HiCA attends on human faces (and persons) when there is speech activity.
- ❑ The VAD performance of HiCA is moderate: accuracy **66.1%**, F score: **55.7%**.

Dataset

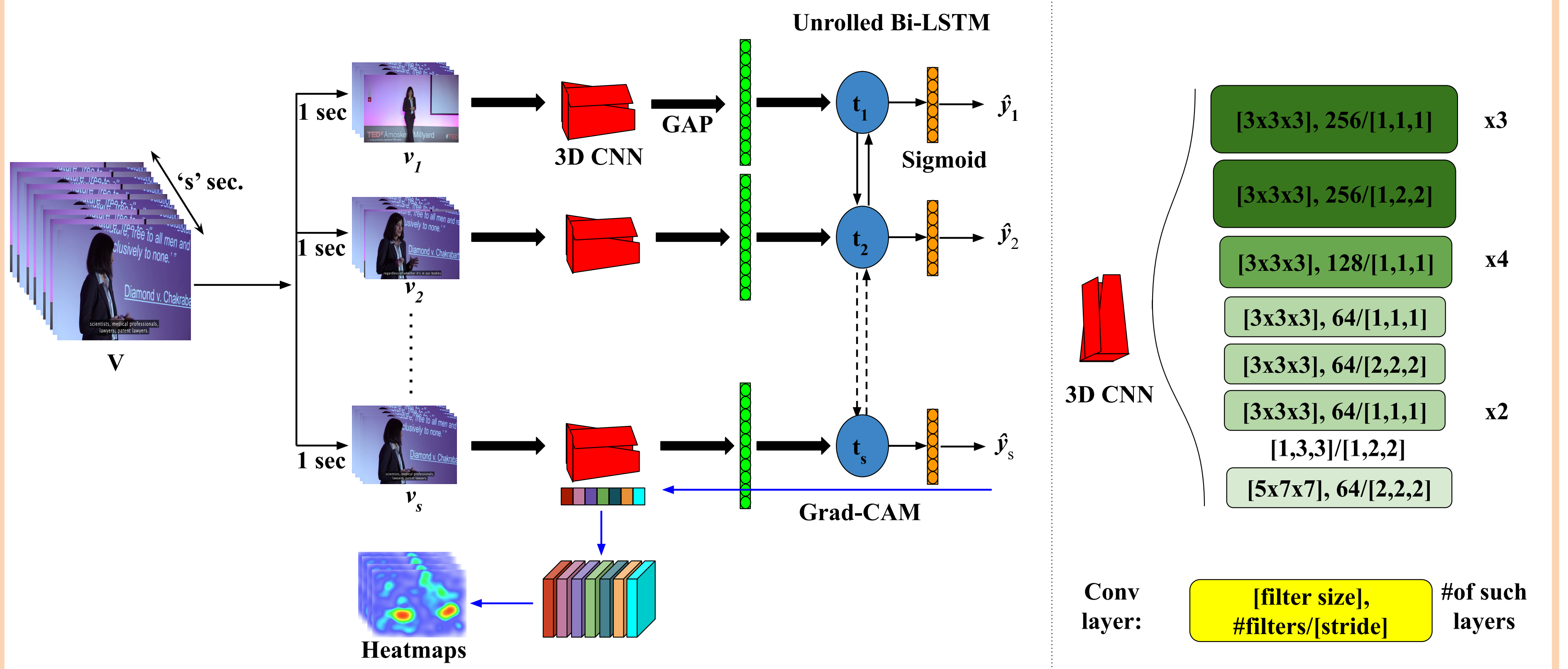
Media content: 97 Hollywood movies + Subtitles

	Speech (hours)	Non-speech (hours)
Training Set	66.64	64.18
Validation Set	16.07	16.28
Test Set	15.86	15.23

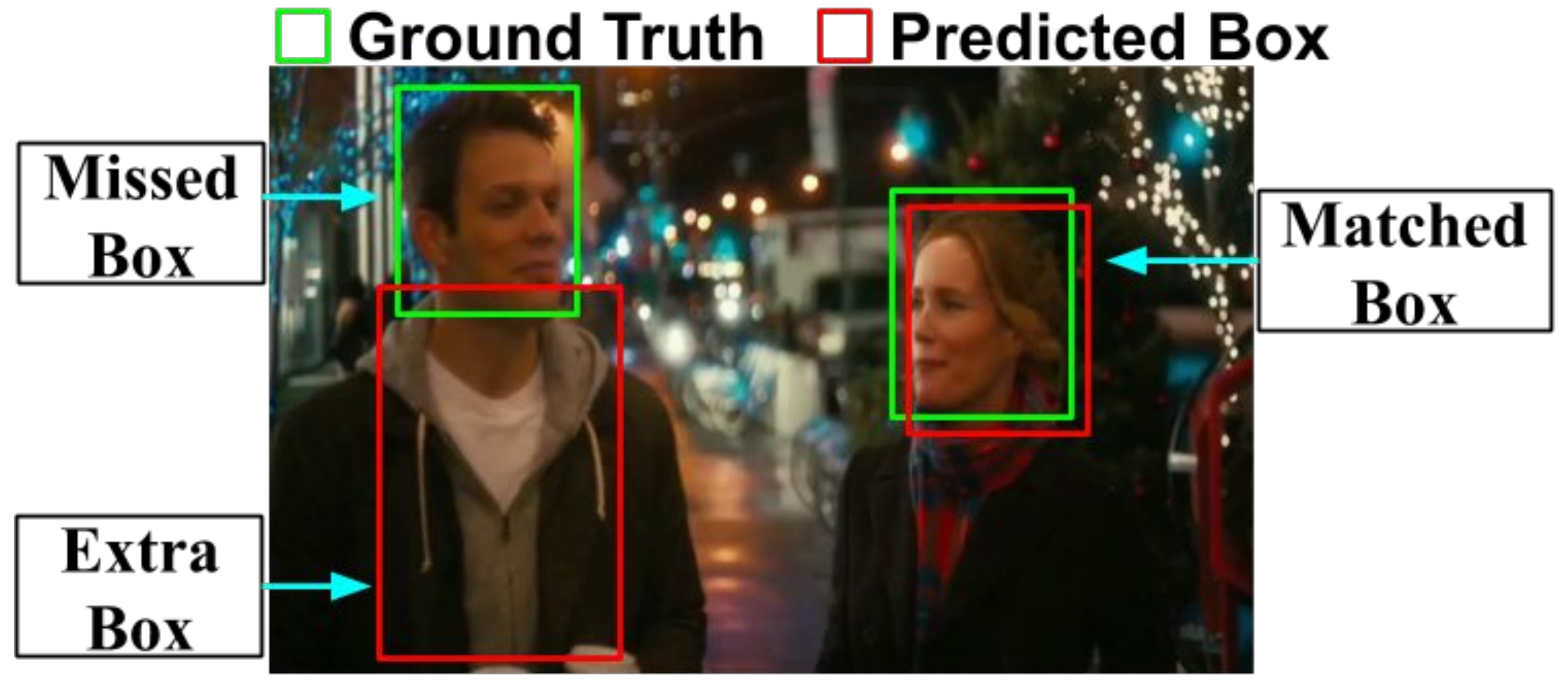
3D Grad-CAMs



Architecture

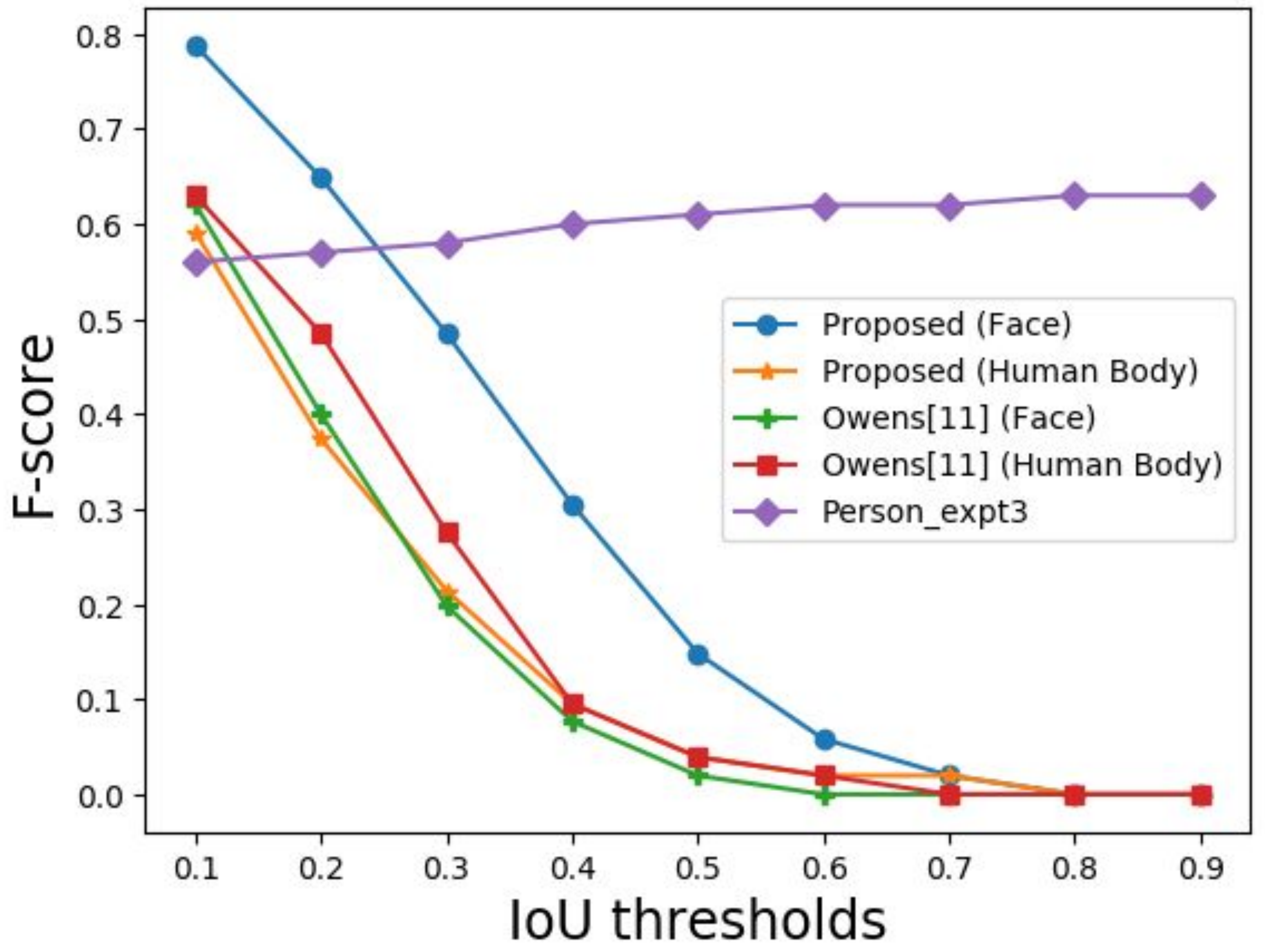


Analysis of CAMs



Example of matched, missed and extra box

Expt1 (Face): Compare against all face detection regions.
Expt2 (Human Body): Compare with all human body proposals.
Expt3 (Person): Analyse non-face predictions.
Baseline: Owens et.al ECCV, 2018.



Trend of F-score for different experiments

Future Work

- ❑ Multimodal fusion to complement audio-VAD systems.
- ❑ Active speaker detection using the learned representations.