Joint Audio-Visual Bi-Modal Codewords for Video Event Detection

Guangnan Ye¹, I-Hong Jhuo², Dong Liu¹, Yu-Gang Jiang³, D. T. Lee²,⁴, Shih-Fu Chang¹

¹Dept. of Electrical Engineering, Columbia University, New York
²Dept. of Computer Science and Information Engineering, National Taiwan University, Taiwan
³School of Computer Science, Fudan University, Shanghai
⁴Dept. of Computer Science and Engineering, National Chung Hsing University, Taiwan

Objective and Overview

Objective: Develop a joint audio-visual bi-modal representation to discover strong audio-visual joint patterns in videos for detecting multimedia events.

- Build a bipartite graph to model relations across the quantized words extracted from the visual and audio modalities;
- Partition the bipartite graph to construct a bi-modal codebook that reveal joint audio-visual patterns;
- Various pooling strategies are employed to re-quantize the visual and audio words into the bi-modal words;
- Bi-modal bag-of-words (BoW) representations are used for event classification.

The Proposed Approach and Experiments

Audio-Visual Bi-Modal BoW Generation

-Nodes are audio words (hₐ) and Visual words (hᵥ).
-Edges denote the correlation of audio and visual words.
-Edge weight (line width) is measured by co-occurrence of audio and visual words, defined by kₐv:

Bipartite Graph Construction

- Various pooling strategies are employed to re-quantize the visual and audio words into the bi-modal words;
- Bi-modal bag-of-words (BoW) representations are used for event classification.

Poolig Strategies

- Average Pooling: \[ h_{av}^k(k) = \frac{\sum_{\sigma \in \mathcal{C}} h_{av}^\sigma(h^p_{av}(\sigma) + h_{av}^\sigma(h^v_{av}(\sigma))}{|\mathcal{C}|} \]
- Max Pooling: \[ h_{max}^k(k) = \max (\sum_{\sigma \in \mathcal{C}} h_{av}^\sigma(h^p_{av}(\sigma), \sum_{\sigma \in \mathcal{C}} h_{av}^\sigma(h^v_{av}(\sigma))) \]
- Hybrid Pooling: \[ h_{hy}^k(k) = \frac{1}{2} (\max (\sum_{\sigma \in \mathcal{C}} h_{av}^\sigma(h^p_{av}(\sigma), \sum_{\sigma \in \mathcal{C}} h_{av}^\sigma(h^v_{av}(\sigma)))) \]

Experiment on TRECVID MED 2011

Effect of varying bi-modal codebook size; average pooling performs the best

Experiment on CCV Dataset

Effect of varying bi-modal codebook size; average pooling performs the best

Summary

- Joint bi-modal codewords achieved 19.6% and 8.6% improvement over LF baseline in TRECVID and CCV, respectively.
- 47% and 36% of bi-modal codewords contain contributions from both modalities in TRECVID and CCV, respectively.
- Among the evaluated pooling strategies, average pooling achieved the best performance.
- Events with multimodal cues, such as “Bird” and “Wedding Ceremony”, show the highest gains.