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

Guangnan Ye [1], I-Hong Jhuo[2], Dong Liu[1], Yu-Gang Jiang[3] , D. T. Lee[2,4], Shih-Fu Chang[1]
[1] Dept.of Electrical Engineering, Columbia University, New York
[2] Dept.of Computer Science and Information Engineering, National Taiwan University, Taiwan
[3] School of Computer Science, Fudan University, Shanghai
[4] Dept. of Computer Science and Engineering, National Chung Hsing University, Taiwan

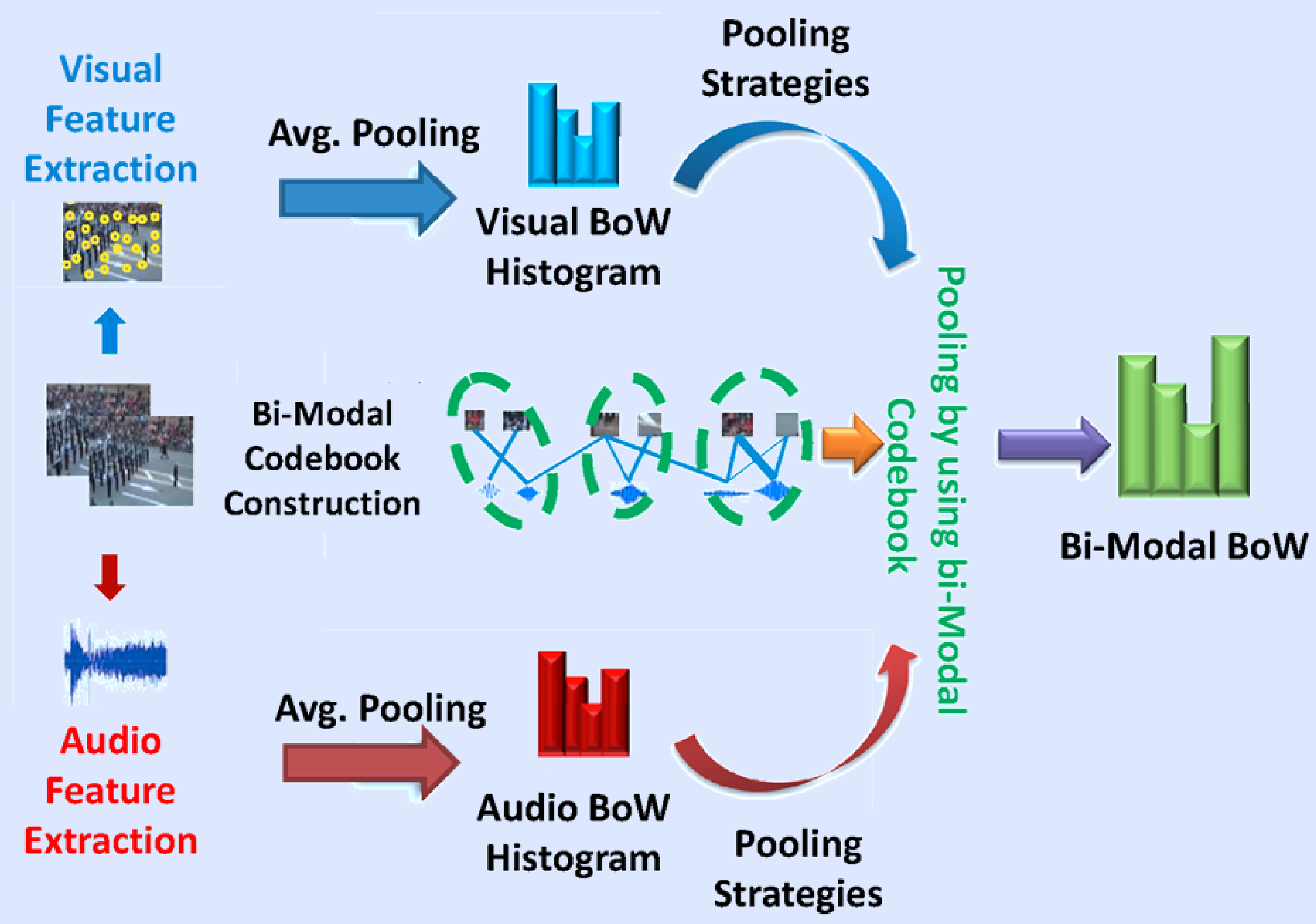
Objective and Overview

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.

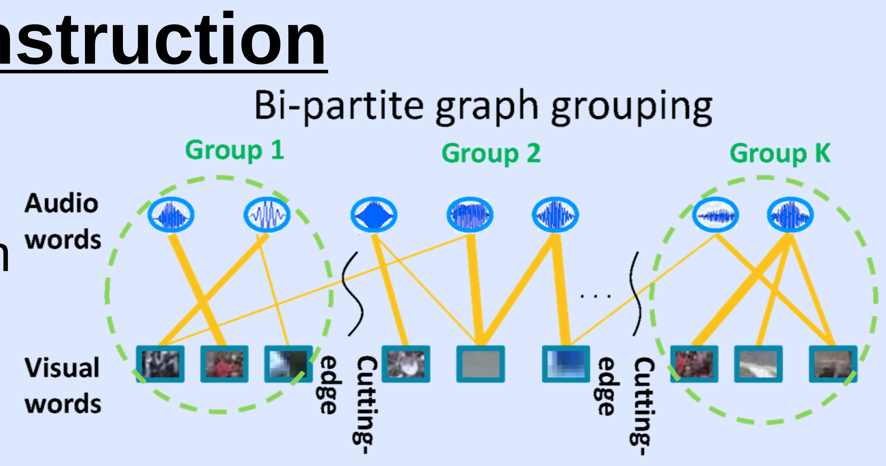
The Proposed Approach and Experiments

Audio-Visual Bi-Modal BoW Generation



Bipartite Graph Construction

- Nodes are audio words (h_i^a) and Visual words (h_i^v).
- 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 s_{kl}:



$$s_{kl} = \frac{\sum_{i=1}^{n} \mathbf{h}_i^a(k)\mathbf{h}_i^v(l)}{\sum_{i=1}^{n} \mathbf{h}_i^a(k) \sum_{i=1}^{n} \mathbf{h}_i^v(l)}$$

The Algorithm

Algorithm 1 Audio-Visual Bi-Modal BoW Representation Generation Procedure

1: **Input:** Training video collection $\mathcal{D} = \{d_i\}$ where each d_i is represented as a multi-modality representation $d = \{\mathbf{h}_i^a, \mathbf{h}_i^v\}$; Size of the audio-visual bi-modal codebook K.
2: Produce the correlation matrix $\mathbf{S}$ between the audio and visual words by calculating the co-occurrence probability over $\mathcal{D}$ by Eq. (1).
3: Calculate matrix $\mathbf{D}_1$, $\mathbf{D}_2$ and $\hat{\mathbf{S}}$ respectively.
4: Apply SVD on $\hat{\mathbf{S}}$ and select $l = \lceil \log_2 K \rceil$ of its left and right singular vectors $\mathbf{U} = [\mathbf{u}_2, \ldots, \mathbf{u}_{l+1}]$ and $\mathbf{V} = [\mathbf{v}_2, \ldots, \mathbf{v}_{l+1}]$.
5: Calculate $\mathbf{Z} = (\mathbf{D}_1^{-1/2}\mathbf{U}, \mathbf{D}_2^{-1/2}\mathbf{V})^\top$.
6: Apply k-means clustering algorithm on $\mathbf{Z}$ to obtain K clusters, which form the audio-visual words $\mathcal{B} = \{B_1, \ldots, B_K\}$.
7: Apply a suitable pooling strategy to re-quantize each video into the audio-visual bi-modal BoW representation.
8: **Output:** Audio-visual BoW representation.

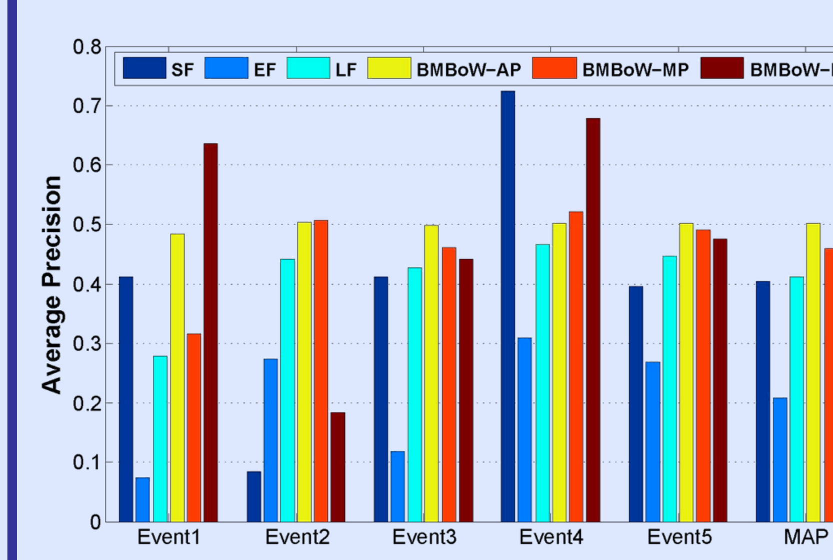
Pooling Strategies

Average Pooling: $$\mathbf{h}_i^{\text{avg}}(k) = \frac{\sum_{w_p^a \in \mathcal{W}_k^a, w_q^v \in \mathcal{W}_k^v} (\mathbf{h}_i^a(p) + \mathbf{h}_i^v(q))}{|\mathcal{W}_k^a| + |\mathcal{W}_k^v|}$$

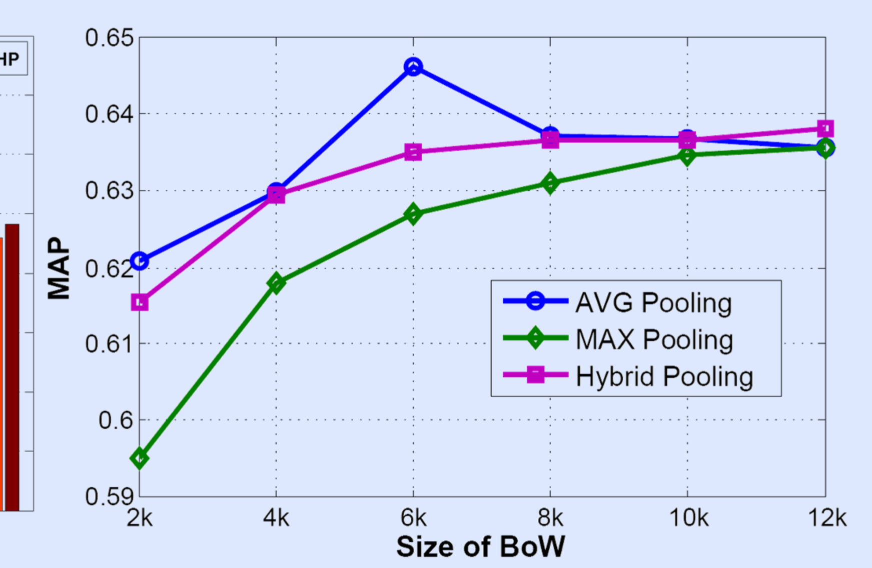
Max Pooling: $$\mathbf{h}_i^{\max}(k) = \max\Big(\sum_{w_p^a \in \mathcal{W}_k^a} \mathbf{h}_i^a(p), \sum_{w_q^v \in \mathcal{W}_k^v} \mathbf{h}_i^v(q)\Big)$$

Hybrid Pooling: $$\mathbf{h}_i^{\text{hyb}}(k) = \frac{1}{2}\Big(\max_{w_p^a \in \mathcal{W}_k^a} \mathbf{h}_i^a(p) + \frac{\sum_{w_q^v \in \mathcal{W}_k^v} \mathbf{h}_i^v(q)}{|\mathcal{W}_k^v|}\Big)$$

Experiment on TRECVID MED 2011

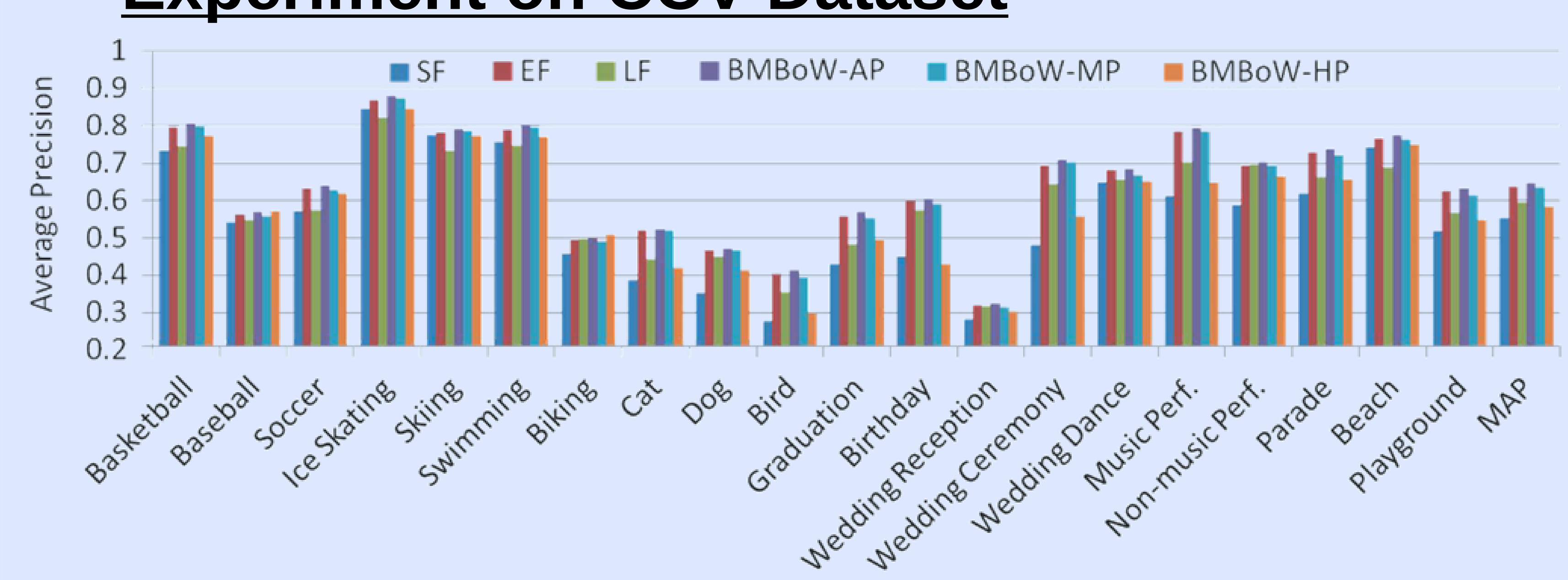


Per-event AP performance, 19.6% gain over LF baseline

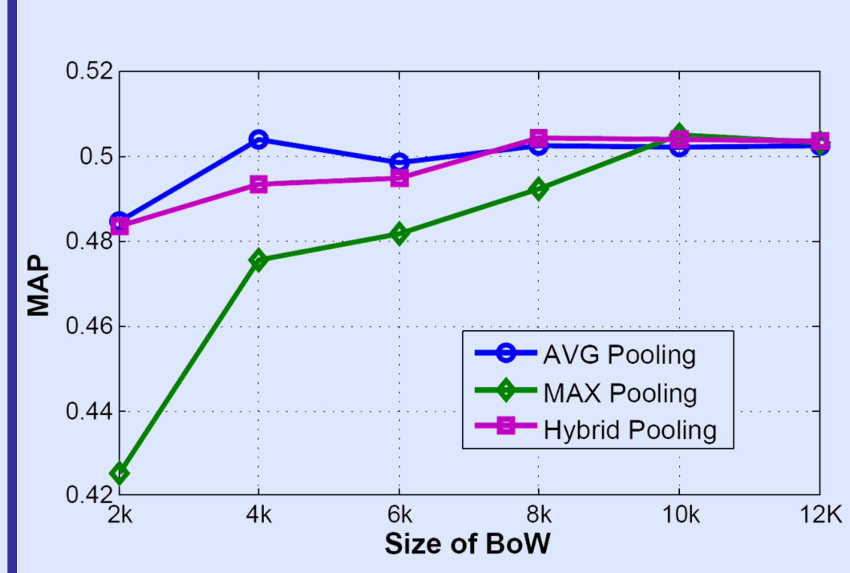


Effect of varying bi-modal codebook size; average pooling performs the best

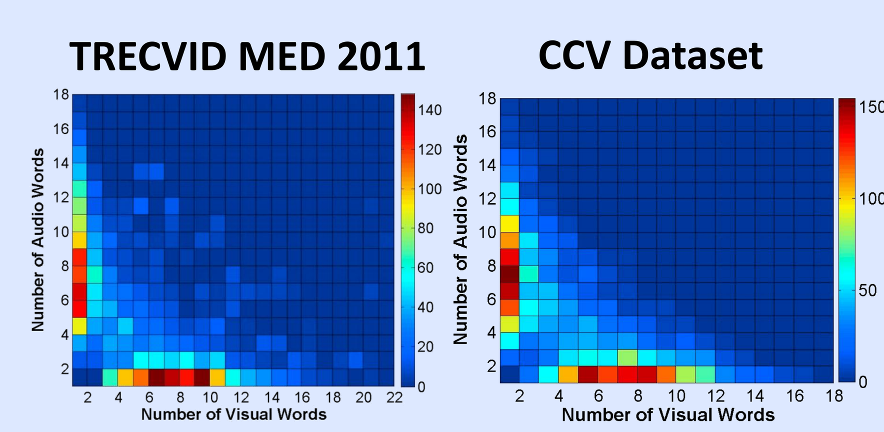
Experiment on CCV Dataset



Per-event AP performance; 8.6% gain over LF baseline



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



Density of audio and visual words within the bi-modal words: 47% and 36% of bi-modal words contain both contributions from audio and visual in TRECVID and CCV dataset respectively.

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.