Machine Vision and Applications manuscript No.
(will be inserted by the editor)

Discovering Joint Audio-Visual Codewords for Video Event Detection

I-Hong Jhuo[1], Guangnan Ye[3]
Shenghua Gao[4], Dong Liu[3],
Yu-Gang Jiang[5],
D. T. Lee[1,2], Shih-Fu Chang[3]

Received: date / Accepted: date

Abstract Detecting complex events in videos is intrinsically a multimodal problem since both audio and visual channels provide important clues. While conventional methods fuse both modalities at a superficial level, in this paper we propose a new representation—called bi-modal words—to explore representative joint audio-visual patterns. We first build a bipartite graph to model relation across the quantized words extracted from the visual and audio modalities. Partitioning over the bipartite graph is then applied to produce the bi-modal words that reveal the joint patterns across modalities. Different pooling strategies are then employed to re-quantize the visual and audio words into the bi-modal words and form bi-modal Bag-of-Words representations. Since it is difficult to predict the suitable number of the bi-modal words, we generate bi-modal words at different levels (i.e., codebooks with different sizes), and use multiple kernel learning to combine the resulted multiple representations during event classifier learning. Experimental results on three popular datasets show that the proposed method achieves statistically significant performance gains over methods using individual visual and audio feature alone and existing popular multi-modal fusion methods. We also find that average pooling is particularly suitable for bi-modal representation, and using multiple kernel learning (MKL) to combine multi-modal representations at various granularities is helpful.

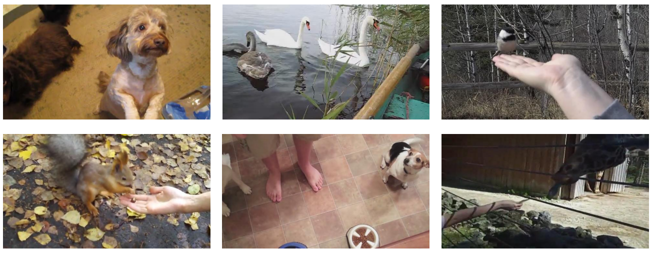


Fig. 1 Example video frames of event "feeding an animal" defined in TRECVID Multimedia Event Detection Task 2011. As can be seen, event detection in such unconstrained videos is a highly challenging task since the content is extremely diverse.

Keywords Bi-Modal Words, Multimodal Fusion, Multiple Kernel Learning, Event Detection.

1 Introduction

Automatically detecting complex events in diverse Internet videos is a topic that is receiving increasing research attention in computer vision and multimedia. Currently a large portion of the Internet videos are captured by amateur consumers without professional post-editing. This makes the task of event recognition extremely challenging, since such videos contain large variations in lighting, viewpoint, camera motion, etc. Figure 1 shows example frames from six videos containing event "feeding an animal". In addition to the variations mentioned above, the "high-level" nature of

[1] Dept. of Computer Science and Information Engineering, National Taiwan University
[2] Dept. of Computer Science and Engineering, National Chung Hsing University
[3] Dept. of Electrical Engineering, Columbia University, New York, USA
[4] Advanced Digital Sciences Center, Singapore
[5] School of Computer Science, Fudan University, Shanghai, China
ihjhuo@gmail.com
shenghua.gao@adsc.com.sg
ygj@fudan.edu.cn
{yegn,dongliu,sfchang}@ee.columbia.edu
dtlee@ieee.org

the event categories (e.g., different kinds of animals in this event) sets a big challenge in event recognition.

Fortunately, besides the visual frames shown in Figure 1, videos also contain audio information which provides an extra useful clue for event detection. In other words, events captured in the videos are multimodal and videos of the same event typically show consistent audio-visual patterns. For example, an "explosion" event is best manifested by the transient burst of sound together with the visible smoke and flame after the incident. Other examples include strong temporal synchronization (e.g., horse running with audible footsteps) or loose association (e.g., people feeding an animal while also talking about the feeding action). Therefore, we believe that successful event detection solutions should effectively harness both audio and visual modalities.

Most existing works fused multimodal features in a superficial fashion, such as early fusion which concatenates feature vectors before classification, or late fusion which combines prediction scores after classification. To better characterize the relationship between audio-visual modalities in videos, in this paper, we propose an audio-visual bi-modal Bag-of-Words (BoW) representation. First, we apply the typical BoW representation to build an audio BoW representation and a visual BoW representation, where the codebooks are generated by using standard k-means clustering separately. After that, a bipartite graph is constructed to capture joint co-occurrence statistics between the quantized audio words and visual words. A bi-modal codebook is then generated by spectral clustering, which partitions the graph into a set of visual/audio word groups, and each group is treated as a joint bi-modal word. Finally, the original individual feature in each modality (audio, or visual) is re-quantized based on the bi-modal codewords, using popular feature pooling methods. In addition, as it is difficult (if not impossible) to predict the suitable number of bi-modal words, we generate bi-modal codebooks of different sizes and employ MKL to combine their respective representations for event model learning. The flowchart of our approach is illustrated in Figure 2.

The main contributions of this paper are summarized as follows:

- We propose the audio-visual bi-modal BoW representation, which effectively explores the underlying structure of the joint audio-visual feature space of complex unconstrained videos. Our representation is very easy to implement because only classical bipartite graph partition technique is used to generate the bi-modal words. Compared with the original audio or visual BoW representations, the joint bi-modal BoW not only outperforms simple early/late fusion, but also greatly reduces the dimensionality of the final video representation.

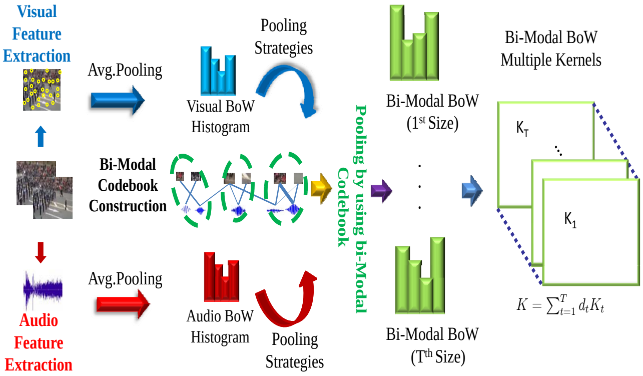


Fig. 2 The framework of our proposed joint bi-modal word representation. We first extract audio and visual features from the videos and then quantize them into audio and visual BoW histograms respectively. After that, a bipartite graph is constructed to model the relations across the quantized words extracted from both modalities, in which each node denotes a visual or an audio word and edges between two nodes encode their correlations. By partitioning the bipartite graph into a number of clusters, we obtain several bi-modal words that reveal the joint audio-visual patterns. With the bi-modal words, the audio and visual features in the original BoW representations are re-quantized into a bi-modal BoW representation. Finally, bi-modal codebooks of various sizes are combined in a multiple kernel learning framework for event model learning.

- Other than fixing the number of codewords as most existing works on visual/audio word-based representations, we propose to generate bi-modal codewords at different granularities (multiple codebooks of different sizes), and adopt MKL [11,17,30] to incorporate multiple bi-modal BoW representations for event detection, which further improves the detection accuracy.

This paper is an extension of a previous conference paper [33] with the new work of using multiple bi-modal codebooks with different granularities and MKL, additional experiments on one more dataset, and extra detailed discussions. The rest of the paper is organized as follows. We first review related works in Section 2. Section 3 discusses typical representations of audio and visual features. Section 4 introduces our proposed audio-visual bi-modal BoW representation. Extensive experimental evaluations on three popular datasets will be given in Section 5. Finally, we conclude this work in Section 6.

2 Related Works

Fusing complementary audio and visual information is important in video content analysis, which has been attempted in many prior works. For example, Jiang *et* *al.* [15] adopted average late fusion, which uses the average prediction scores of multiple independently trained classifiers. Differently, the work in [4] averaged the kernel matrices obtained from audio and visual features before classification, which is known as the early fusion method. Different from these superficial fusion methods, our bi-modal BoW representation characterizes the joint patterns across the two modalities, which is able to uncover their underlying relations rather than simple combination.

There are also several interesting works on joint audio-visual analysis, especially for object tracking and detection. For instance, Beal *et al.* [5] developed a joint probability model of audio and visual cues for object tracking. Cristani *et al.* [8] tried to synchronize foreground objects and foreground audio sounds in the task of object detection. One limitation of these methods is that they only considered videos in a fully controlled environment, which is much easier than the unconstrained videos as handled in this work.

More recently, Jiang *et al.* [12] proposed Short-Time Audio-Visual Atom as the joint audio-visual feature for video concept classification. First, visual regions are tracked within short-term video slices to generate so called visual atoms, and audio energy onsets are located to generate audio atoms. Then the regional visual features extracted from the visual atoms and the spectrogram features extracted from the audio atoms are concatenated to form an audio-visual atom feature representation. Finally, a discriminative joint audio-visual codebook is constructed on the audio-visual atoms using multiple instance learning, and finally the codebook-based BoW features are generated for semantic concept detection. As an extension of this work, in [13], the authors further proposed Audio-Visual Grouplets by exploring temporal audio-visual interactions, where an audio-visual grouplet is defined as a set of audio and visual codewords that are grouped together based on their strong temporal correlations in videos. Specifically, the authors conducted foreground/background separation in both audio and visual channels, and then formed four types of audio-visual grouplets by exploring the mixed-and-matched temporal audio-visual correlations, which provide discriminative audio-visual patterns for classifying semantic concepts. Despite the close relatedness with our work, the above two methods require to perform object or region tracking, which is extremely difficult and computationally expensive, particularly for the unconstrained Internet videos. Several other works demonstrated the success in utilizing audio and visual information for recognition [16,28,31], but are restricted to videos containing emotional music or talking faces. In this paper, our method is proposed for more general situations, and avoids using expensive and unreliable region segmentation and tracking.

Methodologically, our work uses bipartite graph partitioning technique [9] to obtain the bi-modal codebooks. Bipartite graph partitioning has been widely adopted in many applications. For example, Liu *et al.* [21] used a bipartite graph to model the co-occurrence of two related views based on visual vocabularies, and graph partitioning algorithm was applied to find visual word co-clusters. The generated co-clusters not only transfer knowledge across different views but also allow cross-view action recognition. In order to model the co-occurrence relations between words from different domains, Pan *et al.* [26] adopted a bipartite graph and spectral clustering to discover cross-domain word clusters. In this way, the clusters can reduce the gap between different domains, and achieve good performance in cross-domain sentiment classification. In contrast to these applications which focus on cross-domain/view learning, we propose to use a bipartite graph to discover the correlations between audio and visual words. Another algorithm used in our approach is MKL [29, 30], which has been frequently adopted in many computer vision and multimedia tasks.

3 Unimodal Feature Representations

Before introducing the bi-modal BoW representation, let us briefly describe popular unimodal BoW feature representations, which are the basis of our approach. Typical audio/visual BoW representation involves three steps: First, a set of descriptors (visual/audio) are extracted from a video corpus. Then the descriptors are used to generate visual/audio codebooks using k-means clustering. Each cluster describes a common pattern of the descriptors, and is usually referred to as a codeword. With the codebook, feature pooling is performed which aggregates all the descriptors in each video[1] to form a single fixed dimensional feature vector.

We describe the visual/audio descriptors applied in this work as below:

- **Static Sparse SIFT Appearance Feature**. The effectiveness of SIFT descriptors [18] has been proved in numerous object and scene recognition tasks. It is therefore adopted to characterize the static visual information in video frames. Following the work of [15], we adopt two versions of sparse keypoint detector: Difference of Gaussians [18] and Hessian Affine [24], to find local keypoints in the frames.

[1] Normally event detection is performed on video level, i.e., to detect whether a video contains an event of interest. Therefore we represent each video by a feature vector.

Each descriptor for the keypoint is described by a 128-dimensional SIFT vector. To reduce the computational cost, we sample one frame every two seconds. Finally, the SIFT features within a frame are further quantized using a SIFT codebook and form a $5,000$-dimensional BoW histogram.

- **Motion-based STIP Feature**. Motion information is always an important clue for video content recognition. For this, we adopt the commonly used Spatial-Temporal Interest Points (STIP) [20]. STIP extracts space-time local volumes which have significant variations in both space and time. We apply Laptev's algorithm [19] to locate the volumes and compute the corresponding descriptors. Specifically, a local volume is described by the concatenation of Histogram of Gradients (HOG) and Histogram of Optical Flow (HOF). This leads to a 144-dimensional vector for each volume, which is then quantized with a codebook to produce a $5,000$- dimensional BoW histogram.
- **Acoustic MFCC Feature**. Besides the aforementioned visual features, audio information provides another important clue for video event detection [31]. To utilize this, we adopt the popular Mel-Frequency Cepstral Coefficients (MFCC) [27] and compute a 60-dimensional MFCC feature for every temporal window of 32ms. The features are densely computed with nearby windows having 50% overlap. Finally, the MFCC features are quantized into a $4,000$ -dimensional BoW histogram, using the same way as we quantize the visual features.

With these unimodal features, for each video clip, we have a $10,000$-dimensional visual BoW representation by concatenating the BoW histograms generated from SIFT and STIP $(5,000 + 5,000)$, and a $4,000$ dimensional audio BoW representation. These are used to compute the bi-modal representation as discussed in the next section.

4 Joint Audio-Visual Bi-Modal Words

We now introduce the audio-visual bi-modal representation in detail. We first introduce the construction of the bipartite graph based on the audio BoW and the visual BoW representation, and the way of generating the bi-modal codewords. Then we describe three pooling strategies which are used for re-quantizing the original visual/audio BoW into the joint audio-visual bi-modal BoW representation. At the end, we discuss how to integrate the bi-modal BoW representations generated at different granularities using MKL.

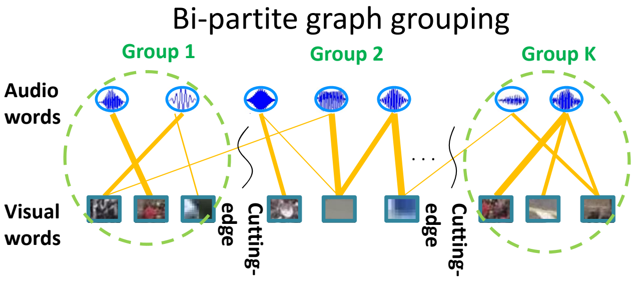


Fig. 3 An illustration of the bipartite graph constructed between audio and visual words, where the upper vertices denote the audio words and the lower vertices denote the visual words. Each edge connects one audio word and one visual word, which is weighted by the correlation measure calculated based on Eq. (1). In this figure, the thickness of the edge reflects the value of the weight.

4.1 Audio-Visual Bipartite Graph Construction

Let $\mathcal{D} = \{d_i\}_{i=1}^n$ be a training collection with n videos. Denote the audio BoW feature of video d_i as $\mathbf{h}_i^a$ and its visual BoW feature as $\mathbf{h}_i^v$, i.e., $d_i = \{\mathbf{h}_i^a, \mathbf{h}_i^v\}$, where $\mathbf{h}_i^a$ is $4,000$-dimensional and $\mathbf{h}_i^v$ is $10,000$-dimensional. These features are ℓ_1 normalized such that the sum of its entries equals to 1. In addition, we use $\mathcal{W}^a = \{w_1^a, \dots, w_{m_a}^a\}$ and $\mathcal{W}^v = \{w_1^v, \dots, w_{m_v}^v\}$ to denote the sets of audio and visual words respectively, where $w_i^a \in \mathcal{W}^a$ represents an audio word and $w_i^v \in \mathcal{W}^v$ indicates a visual word, and m_a and m_v denote the number of audio and visual words respectively. The total number of audio and visual words is $m = m_a + m_v$.

We further define an undirected graph $G = (V, E)$ between the audio and visual words, where V and E denote the set of vertices and the set of edges respectively. Let V be a finite set of vertices $V = V^a \bigcup V^v$, where each vertex in V^a corresponds to an audio word in $\mathcal{W}^a$ and each vertex in V^v corresponds to a visual word in $\mathcal{W}^v$. An edge in E connects two vertices in V^a and V^v, and there is no intra-set edge connecting two vertices in V^a or V^v respectively. This graph $G = (V, E)$, where $V = V^a \bigcup V^v$, is commonly called a *bipartite* graph. To measure the correlation between an audio word $w_k^a \in \mathcal{W}^a$ and a visual word $w_l^v \in \mathcal{W}^v$, we assign a non-negative weight s_{kl} to any edge $e_{kl} \in E$, which is defined as follows,

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

where $\mathbf{h}_i^a(k)$ denotes the entry of $\mathbf{h}_i^a$ corresponding to the kth audio word w_k^a and $\mathbf{h}_i^v(l)$ denotes the entry of $\mathbf{h}_i^v$ corresponding to the lth visual word w_l^v.

In Eq. (1), the numerator measures the summation of the joint probability of the audio word w_k^a and the visual word w_l^v, where the summation is calculated over the entire video collection. This value essentially reveals the correlation of the audio and visual words. On the other hand, the denominator acts as a normalization term, which penalizes the audio and/or visual words that frequently appear in the video collection. It is also worth noting that the choice of the correlation measure in Eq. (1) is flexible. We can also estimate the weight s_{kl} by applying other methods like Pointwise Mutual Information (PMI) [21]. Figure 3 gives a conceptual illustration of a bipartite graph constructed from the joint statistics of the audio and visual words.

4.2 Discovering Bi-Modal Words

We adopt standard bipartite graph partitioning method to discover the audio-visual bi-modal words. Following [9], we begin with a bipartitioning method over the bipartite graph and then extend it into the multipartitioning scenario.

Recall that we have a bipartite graph $G = (V, E)$ between the audio and visual words. Given a partitioning of the vertex set V into two subsets V_1 and V_2, the cut between them can be defined as sum of all edge weights connecting vertices from the two subsets,

$$\text{cut}(V_1, V_2) = \sum_{k \in V_1, l \in V_2} s_{kl}. \tag{2}$$

The bipartite partition problem over the bipartite graph is to find the vertex subsets V_1^* and V_2^* such that $\text{cut}(V_1^*, V_2^*) = \min_{V_1, V_2} \text{cut}(V_1, V_2)$. To this end, we define the Laplacian matrix $\mathbf{L} \in \mathbb{R}^{m \times m}$ associated with the bipartite graph G as,

$$L_{kl} = \begin{cases} \sum_l s_{kl}, & k = l, \\ -s_{kl}, & k \neq l \text{ and } e_{kl} \in E, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Given a bipartitioning of V into V_1 and V_2, we further define a partition vector $\mathbf{p} \in \mathbb{R}^m$ that characterizes this division, in which the ith entry describes the partitioning state of $i \in V$,

$$p_i = \begin{cases} +1, & i \in V_1, \\ -1, & i \in V_2. \end{cases} \tag{4}$$

With the above definitions, it can be proved that the graph cut can be equally written as the following form,

$$\text{cut}(V_1, V_2) = \frac{1}{4}\mathbf{p}^\top \mathbf{L}\mathbf{p} = \frac{1}{4} \sum_{(i,j) \in E} s_{ij}(p_i - p_j)^2. \tag{5}$$

However, it can be easily seen from Eq. (5) that the cut is minimized by a trivial solution when all p_i's are either $+1$ or -1. To avoid this problem, a new objective function is used to achieve not only minimized cut but also a balanced partition. Formally, the objective function is defined as follows,

$$Q(V_1, V_2) = \frac{\text{cut}(V_1, V_2)}{\text{weight}(V_1)} + \frac{\text{cut}(V_1, V_2)}{\text{weight}(V_2)}, \tag{6}$$

where $\text{weight}(V_i) = \sum_{k,l \in V_i} s_{kl}$, $i = 1, 2$. Then it can be proved that the eigenvector corresponding to the second smallest eigenvalue of the generalized eigenvalue problem $\mathbf{Lz} = \lambda \mathbf{Dz}$ (where $\mathbf{D}$ is a diagonal matrix with $D(k,k) = \sum_l s_{kl}$) provides a real relaxed solution of the discrete optimization problem in Eq. (6) [22]. To obtain the the eigenvector corresponding to the second smallest eigenvalue, [9] proposes a computationally efficient solution through Singular Value Decomposition (SVD). Specifically, for the given bipartite graph G, we have

$$\mathbf{L} = \begin{pmatrix} \mathbf{D}_1 & -\mathbf{S} \\ -\mathbf{S}^\top & \mathbf{D}_2 \end{pmatrix}, \text{and } \mathbf{D} = \begin{pmatrix} \mathbf{D}_1 & 0 \\ 0 & \mathbf{D}_2 \end{pmatrix}, \tag{7}$$

where $\mathbf{S} = [s_{kl}]$, $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices such that $D_1(k,k) = \sum_l s_{kl}$ and $D_2(l,l) = \sum_k s_{kl}$. Let the normalized matrix $\hat{\mathbf{S}} = \mathbf{D}_1^{-1/2}\mathbf{S}\mathbf{D}_2^{-1/2}$, it can be proved that the eigenvector corresponding to the second smallest eigenvalue of $\mathbf{L}$ can be expressed in terms of the left and right singular vectors corresponding to the second largest singular value of $\hat{\mathbf{S}}$ as follows,

$$\mathbf{z}_2 = \begin{bmatrix} \mathbf{D}_1^{-1/2}\mathbf{u}_2 \\ \mathbf{D}_2^{-1/2}\mathbf{v}_2 \end{bmatrix}, \tag{8}$$

where $\mathbf{z}_2$ is the eigenvector corresponding to the second smallest eigenvalue of $\mathbf{L}$, $\mathbf{u}_2$ and $\mathbf{v}_2$ are the left and right singular vectors corresponding to the second largest singular value of $\hat{\mathbf{S}}$.

Finally, we need to use $\mathbf{z}_2$ to find the approximated optimal bipartitioning by assigning each $\mathbf{z}_2(i)$ to the clusters $\mathcal{C}_j$ $(j = 1, 2)$ such that the following sum-of-squares criterion is minimized,

$$\sum_{j=1}^{2} \sum_{\mathbf{z}_2(i) \in \mathcal{C}_j} (\mathbf{z}_2(i) - m_j)^2, \tag{9}$$

where m_j is the cluster center of $\mathcal{C}_j$ $(j = 1, 2)$.

The above objective function can be practically minimized by directly applying the k-means clustering method on the 1-dimensional entries of $\mathbf{z}_2$. The bipartitioning method can be easily extended to a general case of finding K audio-visual clusters [9]. Suppose we have $l = \lceil \log_2 K \rceil$ singular vectors $\mathbf{u}_2, \mathbf{u}_3, \ldots, \mathbf{u}_{l+1}$, and $\mathbf{v}_2, \mathbf{v}_3,$

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: Create 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.
```

$\ldots, \mathbf{v}_{l+1}$, then we can form the following matrix with l columns,

$$\mathbf{Z} = \begin{bmatrix} \mathbf{D}_1^{-1/2}\mathbf{U} \\ \mathbf{D}_2^{-1/2}\mathbf{V} \end{bmatrix}, \tag{10}$$

where $\mathbf{U} = [\mathbf{u}_2, \ldots, \mathbf{u}_{l+1}]$ and $\mathbf{V} = [\mathbf{v}_2, \ldots, \mathbf{v}_{l+1}]$. Based on the obtained matrix $\mathbf{Z}$, we further run k-means method on it to obtain K clusters of audio-visual words, which can be represented as follows,

$$\mathcal{B} = \{B_1, \ldots, B_K\}, \tag{11}$$

where each B_i consists of the audio word subset $\mathcal{W}_i^a$ and the visual word subset $\mathcal{W}_i^v$ falling in the same bi-modal cluster. Note that either $\mathcal{W}_i^a$ or $\mathcal{W}_i^v$ can be empty, indicating that only one modality forms a consistent pattern within the bi-modal word B_i (e.g., visual words corresponding to the background scene).

The above graph partition method needs to compute eigenvectors of the Laplacian matrix, and thus has a computational complexity of $\mathcal{O}(m^3)$ in general, where m is the total number of audio and visual words. We implement the method using Matlab with a Six-Core Intel Xeon Processor X5660 (2.8 GHz) and 32 GB memory. It takes 32 minutes to group $14,000$ audio and visual words into $2,000$ bi-modal words in the experiment on the CCV dataset (cf. Section 5.1).

4.3 Bi-Modal BoW Generation

After generating the bi-modal codewords, we need to map the original visual and audio descriptors to the new codebook. The main purpose here is to fuse the original two visual and audio representations into one joint representation, which will be used for event classification. For this, we adopt three different quantization strategies. Given a video $d_i = (\mathbf{h}_i^a, \mathbf{h}_i^v)$, the audio-visual bi-modal BoW representations generated by average pooling, max pooling and hybrid pooling are described as follows.

4.3.1 Average Pooling

Average pooling treats the audio and visual words equally important. Formally, the bi-modal BoW generation strategy is described as follows,

$$\mathbf{h}_i^{\text{avg}}(k) = \frac{\sum_{w_p^a \in \mathcal{W}_k^a, w_q^v \in \mathcal{W}_k^v} (\mathbf{h}_i^a(p) + \mathbf{h}_i^v(q))}{|\mathcal{W}_k^a| + |\mathcal{W}_k^v|}, \tag{12}$$

where w_p^a means the pth audio word, w_q^v represents the qth visual word, and $\mathbf{h}_i^{\text{avg}}(k)$ denotes the entry in the bi-modal BoW $\mathbf{h}^{avg}$ corresponding to a given audio-visual bi-modal word $B_k = (\mathcal{W}_k^a, \mathcal{W}_k^v)$. $|\mathcal{W}_k^a|$ and $|\mathcal{W}_k^v|$ denote the cardinalities of $\mathcal{W}_k^a$ and $\mathcal{W}_k^v$ respectively. As we can see in Eq.(12), the measure of the entry in the bi-modal representation is the average value of the entries of the audio and visual words in the original BoW representations. We call such bi-modal BoW generation strategy *average pooling* due to its relatedness w.r.t the pooling strategy in sparse coding [6].

4.3.2 Max Pooling

Max pooling selects the maximum summation in the original audio or visual words as the quantization value of the given audio-visual bi-modal word, which is formally defined as follows,

$$\mathbf{h}_i^{\max}(k) = \max\big(\sum_{w_p^a \in \mathcal{W}_k^a} \mathbf{h}_i^a(p), \sum_{w_q^v \in \mathcal{W}_k^v} \mathbf{h}_i^v(q)\big). \tag{13}$$

4.3.3 Hybrid Pooling

We also propose a hybrid pooling strategy which integrates average pooling and max pooling together. Intuitively, the visual features from the visual scene in the video tend to persist over a certain interval when the camera does not move too fast. Therefore, we use average pooling to aggregate information in the interval. Max pooling is employed for the audio information since audio features tends to be transient in time. Formally, the hybrid pooling strategy can be defined as follows,

$$\mathbf{h}_i^{\text{hyb}}(k) = \frac{1}{2}\big(\max_{w_p^a \in \mathcal{W}_k^a} \mathbf{h}_i^a(p) + \frac{\sum_{w_q^v \in \mathcal{W}_k^v} \mathbf{h}_i^v(q)}{|\mathcal{W}_k^v|}\big), \tag{14}$$

where the average pooling aggregates the two entries of the audio and visual words obtained from max and average pooling respectively.

Algorithm 1 gives the detailed flow of the generation procedure of the bi-modal BoW representation.

4.4 Combining Multiple Joint Bi-Modal Representations

As it is true for any BoW-based representations, it is extremely difficult (if not impossible) to identify the suitable number of codewords. Existing works mostly set a fixed number (a few thousands), which have been empirically observed to work well in practice. Since our bi-modal words are generated on top of the audio and visual words, the problem becomes more complicated since there is not (even empirical) evidence of a suitable word number of the joint codebook. Using a small bi-modal codebook will result in ambiguous audio-visual patterns within a bi-modal word. On the other hand, the joint audio-visual patterns may be separated immensely if the codebook size is too large.

To alleviate the effect of codebook size, in this paper, we propose to generate the bi-modal BoW representation at different granularities, i.e., with different codebook sizes. Representations are then combined through the well-known MKL framework. Specifically, suppose we have the joint bi-modal BoW representations generated from T bipartite graph partitioning with different resolutions (i.e., cluster number), and denote the kernel matrix corresponding to the histogram generated at the tth resolution as $K_t(h, h')$. MKL seeks an optimal combination $K(h, h') = \sum_{t=1}^{T} d_t K_t(h, h')$ with the constraints $d_t \geq 0$, $\forall t$ and $\sum_{t=1}^{T} d_t = 1$. By using this $K(h, h')$ for event classification, the performance can be usually boosted as compared with that using a single kernel. Lots of MKL frameworks [11,17, 30] have been proposed and demonstrated for visual classification. In this paper, we adopt the widely used simpleMKL framework [29] because of its sound performance and efficiency. In this MKL framework, for each kernel K_t, it is being associated with a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_t$, and the decision function is in the form $f(h) + b = \sum_t f_t(h) + b$ where each f_t is associated with $\mathcal{H}_t$. The objective of simpleMKL is given as follows:

$$\begin{aligned} \min_{f_t, b, \xi, d} \quad & \frac{1}{2}\sum_t \frac{1}{d} \|f_t\|^2_{\mathcal{H}_t} + C \sum_i \xi_i \\ s.t. \quad & y_i \sum_i f_t(x_i) + y_i b \geq 1 - \xi_i, \quad \forall i \\ & \xi_i \geq 0, \quad \forall i \\ & \sum_t d_t = 1, \quad d_t >= 0, \quad \forall t, \end{aligned} \tag{15}$$

To solve the above objective, we use the simpleMKL solver [29].

5 Experiments

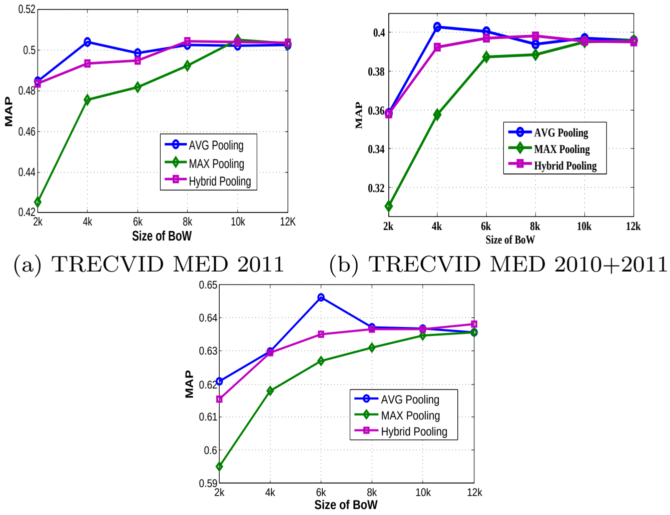
In this section, we evaluate our proposed audio-visual bi-modal representation for video event detection using three datasets: TRECVID Multimedia Event Detection (MED) 2011 dataset, a large dataset that consists of both TRECVID MED 2010 and TRECVID MED 2011, and the recently released Columbia Consumer Video (CCV) dataset.

5.1 Datasets

TRECVID MED 2011 Dataset. TRECVID MED [3] is a challenging task for the detection of complicated high-level events in unconstrained Internet videos. Our first dataset is the MED 2011 development set, which includes five events “Attempting a board trick”, “Feeding an animal”, “Landing a fish”, “Wedding ceremony”, and “Working on a woodworking project”. This dataset consists of $10,804$ videos from $17,566$ minutes of web videos, which is partitioned into a training set ($8,783$ videos) and a test set ($2,021$ videos). The training set contains about 100 positive videos for each event (Most videos in the training set are background videos which do not contain any of the five events). Within each class, there exist complicated content variations, making the task extremely challenging.

TRECVID MED 2010+2011 Dataset. We also consider the earlier MED 2010 dataset [2]. Since the MED 2010 dataset is too small (less than 2,000 videos), we combine TRECVID MED 2010 with MED 2011 [2,3] together to form a larger and more challenging event detection dataset. MED 2010 has three events “Assembling a shelter”, “Batting a run in”, and “Making a cake”, each having 50 positive videos for training and 50 for testing. This combined dataset consists of $14,272$ videos falling into 8 event categories, which is partitioned into a training set ($10,527$ videos) and a test set ($3,745$ videos). Note that the combination also gives another opportunity to re-examine the performance of the five MED 2011 events when more background videos are added (i.e., the MED 2010 videos).

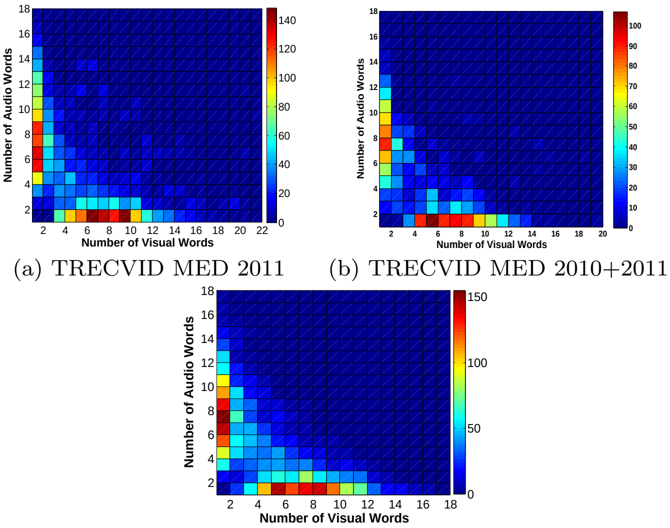
CCV Dataset. CCV dataset [14] contains $9,317$ YouTube videos annotated over 20 semantic categories, where $4,659$ videos are used for training and the remaining $4,658$ videos are used for testing. Most of the 20 categories are events, with a few categories belonging to objects or scenes. To facilitate benchmark comparison, we report performance of all the 20 categories.



(a) TRECVID MED 2011 (b) TRECVID MED 2010+2011

(c) CCV Dataset

Fig. 4 Effect of bi-modal codebook size and pooling strategies on the three datasets.



(a) TRECVID MED 2011 (b) TRECVID MED 2010+2011

(c) CCV Dataset

Fig. 5 The density of audio and visual words in the bi-modal words.

5.2 Experimental Setup

As discussed earlier, we adopt the SIFT BoW (5,000 dimensions) and STIP BoW (5,000 dimensions) representations as the visual features while using the MFCC BoW (4,000 dimensions) as the audio representation. One-vs-all SVM is used to train a classifier for each evaluated event. To get the optimal SVM tradeoff parameter for each method, we partition the training set into 10 subsets and then perform 10-fold cross validation. Moreover, we adopt the χ^2 kernel due to its outstanding performances in many BoW-based applications, which is calculated as $k(x,y) = exp(-\frac{d_{\chi^2(x,y)}}{\sigma})$, where σ is following the previous work [33], $d_{\chi^2(x,y)}$ is defined as $d_{\chi^2(x,y)} = \sum_{i=1} \frac{(x(i)-y(i))^2}{x(i)+y(i)}$, and σ is by default set as the mean value of all pairwise distances in the training set.

For performance evaluation, we follow previous works [14, 25] and use Average Precision (AP), which approximates the area under a precision-recall curve. We calculate AP for each event and then use Mean Average Precision (MAP) across all the event categories in each dataset as the final evaluation metric.

In the following experiments, we will systematically evaluate the performances of the following methods:

1. Single Feature (SF). We only report the best performance achieved by one of the three audio/visual features.
2. Early Fusion (EF). We concatenate the three kinds of BoW features into a long vector with $14,000$ dimensions.
3. Late Fusion (LF). We use each feature to train an independent classifier and then average the output scores of the three classifiers as the final fusion score for event detection.
4. Average Pooling based Bi-Modal BoW (BMBoW-AP), where the average pooling is employed to generate the bi-modal BoW.
5. Max Pooling based Bi-Modal BoW (BMBoW-MP), where we use max pooling to generate the bi-modal BoW.
6. Hybrid Pooling based Bi-Modal BoW (BMBoW-HP), which applies the hybrid pooling to generate the bi-modal BoW.
7. MKL based Bi-Modal BoW (BMBoW-MKL), which uses MKL to combine multiple bi-modal BoW representations. We used all the codebook sizes as experimented in Figure 4.

5.3 The Effect of Codebook Size and Pooling Strategies

We first experimentally evaluate the performance of different codebook sizes and pooling strategies. Since the sizes of audio and visual modalities are 4,000 and 10,000, we expect each bipartite partitioning can provide a high correlation between audio and visual modalities. Therefore, we increase the size of the bi-modal codebook from 2,000 to 12,000 and discuss the MAP performances with different pooling strategies in Figure 4. We can see that average pooling tends to show

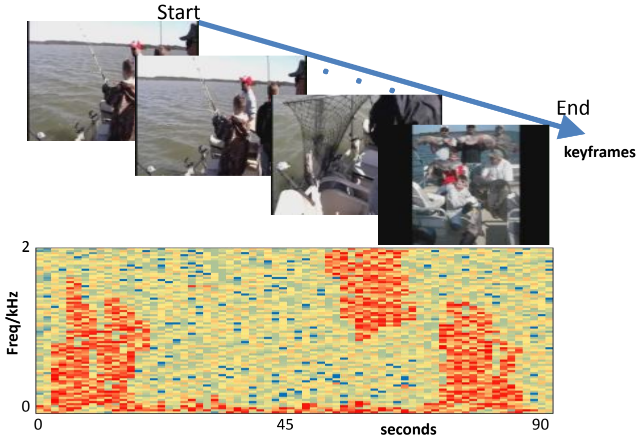


Fig. 6 An example of audio-visual correlations in the event "Landing a fish" from the TRECVID MED 2011 dataset. We see that there are clear audio patterns correlating with the beginning and the end (fish successfully landed) of the event.

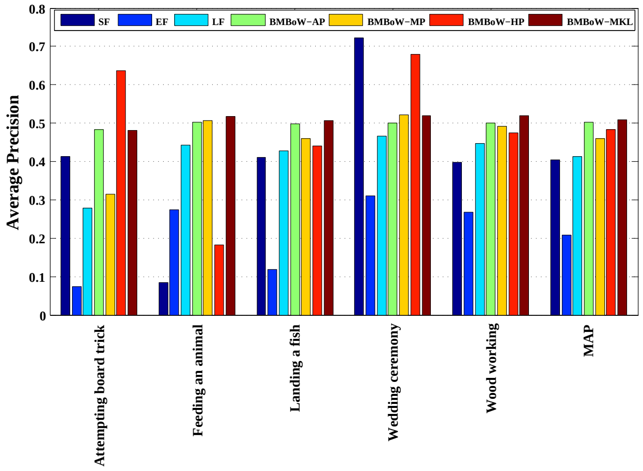


Fig. 7 Per-event performance on TRECVID MED 2011 dataset. This figure is best viewed in color.

better stability than max pooling and hybrid pooling when the codebook size varies, which demonstrates that average pooling is more suitable for the bi-modal BoW quantization. This may be due to the fact that average pooling captures the joint audio-visual patterns while hybrid/max pooling incurs significant information loss caused by considering only the max response of audio/visual information. For codebook size, 4,000 seems to be a good number for the MED datasets, but for CCV, large codebooks with 6,000-10,000 bi-modal words seem to be more effective. Note that for such a large bi-modal codebook, there are codewords containing only word from the audio or visual channel. It makes sense to have such words because, while we would like to discover the correlations between the two modalities, not all the words are correlated. Therefore it is good to leave some audio/visual words independent in the bi-modal representation. This inconsistent observation also confirms the usefulness of aggregating multiple bi-modal codebooks, which will be evaluated later.

We also show the density of audio and visual words within each bi-modal word in Figure 5. Here each grid in the map denotes the frequency of bi-modal words made up of a certain numbers of audio word (vertical coordinate) and visual word (horizontal coordinate). It estimates the portion of the words in the entire bi-modal codebook that contain both visual and audio information, which is found to be about 47% for the TRECVID MED 2011 dataset, 39% on the combined TRECVID MED 2010 and 2011 dataset, and 36% for the CCV dataset. This confirms the significant effect of the audio-visual correlations in the joint bi-modal representation. Therefore, The bi-modal word is also an important component of a large event detection system that achieved the best performance [25] in TRECVID MED 2011. We observe that a large number of bi-modal words contain more visual words than audio words, or the opposite of having more audio words than visual words (i.e., the bins close to x or y axis in Figure 5). This may be due to the fact that some large visual or audio patterns are only correlated to a small clue in the other modality. For instance, a birthday scene with many visual characteristics may be only highly correlated to cheering sounds.

5.4 Performance Comparison on TRECVID MED 2011 Dataset

Let us now evaluate the seven methods listed in Section 5.2. Figure 7 shows results on the MED 2011 dataset. We fix the size of the bi-modal codebook to be 4,000 except for the BMBoW-MKL method, which combines multiple codebook sizes as listed in Figure 4. Also, we adopt average pooling in BMBoW-MKL, since—as will be shown—it outperforms max pooling and hybrid pooling. Based on the results, we have the following findings:

1. Our proposed bi-modal word representation outperforms all the other baseline methods in terms of MAP, which proves the effectiveness of this approach. In particular, it outperforms the most popularly used early fusion and late fusion methods by a large margin. This is due to the fact that the bi-modal words not only capture the correlation between audio and visual information but also aggregate their mutual dependence.
2. As an important but quite obvious conclusion, the bi-modal word representation performs significantly better than all the single features, which verifies the merits of considering multi-modality in the task of video event detection.

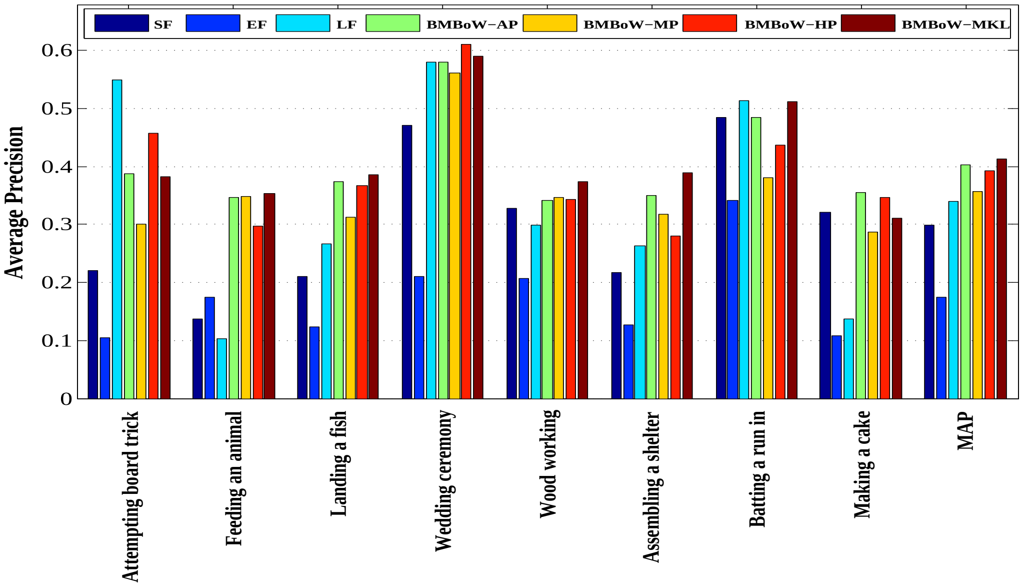


Fig. 8 Per-event performance comparison on TRECVID MED 2010+2011 dataset, which includes eight events. This figure is best viewed in color.

3. As indicated in the introduction of this paper, visual and audio information of the same event category often present consistent joint patterns. This not only holds for events with intuitive audio-visual correlations like “Batting a run in”, but is also true for events that contain fewer audio clues. Figure 6 gives an example. In the event “Landing a fish”, although the soundtrack is mostly quite silent, in the beginning and after the fish is successfully landed, there are some clear audio patterns. Our method is able to capture such local correlations, which is the main reason that it performs better than the simple fusion strategies.
4. BMBoW-AP tends to give better results that BMBoW-MP, which may be due to the fact that the former captures the joint audio-visual patterns while the latter incurs significant information loss caused by selecting only the max contribution between two modalities.
5. BMBoW-HP outperforms BMBoW-MP, as it utilizes perhaps the more suitable pooling strategies for different modalities (i.e., max pooling for the transient audio signal and average pooling for the persistent visual signal). For some events, BMBoW-HP even achieves better results than BMBoW-AP, indicating that selecting maximum response of audio signal may help reveal the semantic clue of the videos. However, in general BMBoW-AP is the best among the three pooling strategies.
6. BMBoW-MKL shows better results than all the methods based on a single bi-modal codebook, confirming the fact that using multiple codebooks is helpful.

5.5 Performance Comparison on TRECVID MED 2010+2011 Dataset

Figure 8 shows the per-event performance for all the methods on this combined dataset. From the results, we can see that the MKL based bi-modal representation, i.e., BMBoW-MKL, achieves the best performance. Specifically, it outperforms the BMBoW-AP, BMBoW-MP, and BMBoW-HP by 0.96%, 5.53%, and 1.96% respectively in terms of MAP. Among the three pooling methods, average pooling offers the best result. In addition, comparing results of the five 2011 events on this combined dataset with that on MED 2011, we also observe that the performances of early and late fusion are not as stable as that of the bi-modal representations when more background videos are added. For example, late fusion performs quite badly for “attempting board trick” event on MED 2011 but is very good on the combined dataset.

5.6 Performance Discussion on CCV Dataset

Figure 9 further shows the per-category performance comparison on the CCV dataset, where the bi-modal codebook size is set as 6,000, except the BMBoW-MKL method. Again, the results show that the BMBoW-MKL achieves the best performance in terms of MAP. It outperforms the BMBoW-AP, BMBoW-MP and BMBoW-HP by 1.16%, 2.26% and 7.36% respectively. Moreover, BMBoW-MKL also achieves the best performance on most of the event categories. For instance, on event “graduation”, it outperforms the best baseline method

SF by 15.05%. Besides, comparing with the best baseline EF, our method achieves the highest relative performance gain on category "birds " and "wedding ceremony". This may be because these two categories contain more significant audio-visual correlations than the other categories. For example, the appearance of birds is often accompanied with the singing sound. Meanwhile, people's actions in a wedding ceremony are always accompanied by background music. In general, we expect high impact of the proposed bi-modal features on other events that share strong audio-visual correlations like the ones mentioned above.

5.7 Statistical Significance Testing

We also measure the statistical significance between the best baseline and BMBoW-MKL on the three datasets. We use a popular measure for statistical significance testing, the p-value, which is the probability of obtaining a test statistic at least as extreme as the one that was observed, assuming that the null hypothesis is true [1]. We can reject the null hypothesis when the p-value is less than the significance level, which is often set as 0.05. When the null hypothesis is rejected, the result is said to be statistical significant. In order to get the p-value, we sample 50% of the test set from each dataset and repeat the experiment 1000 times. For each round, we compute the paired MAP differences $D_i = MAP_{\mathrm{BMBoW-MKL}}(i) - MAP_{\mathrm{Baseline}}(i)$, where $i = 1, 2, \ldots, 1000$. Then we make the assumption that the null hypothesis is $D_i < 0, i = 1, 2, \ldots, 1000$, based on which, the p-value can be defined as the percentage of D_i that is below 0. We find that the p-values obtained on the MED 2011, MED 2010+2011, and CCV datasets are 0.015, 0.018 and 0.022 respectively, which are well below 0.05 and show that the null hypothesis can be rejected. Therefore, we can conclude that our method has achieved statistical significant improvements over the best baseline on the three datasets.

6 Conclusion

In this paper, we have introduced a bi-modal representation to better explore the power of audio-visual features in video event detection. The proposed method uses a bi-partite graph to model the relationship between visual and audio words and partitions the graph to generate audio-visual bi-modal words. Several popular pooling methods are evaluated to generate the BoW representation using the bi-modal words, and average pooling is found to be the best performer. Extensive experiments on three datasets consistently show that the proposed bi-modal representation significantly outperforms early and late fusion, which are currently the most widely used multimodal fusion methods. In addition, since there is no single codebook size that is suitable in all cases, we propose to use multiple bi-modal codebooks and MKL to combine BoW representations based on different codebooks. Results show that using MKL and multiple bi-modal codebooks is always helpful. With these findings we conclude that many state-of-the-art video event detection systems may have overlooked the importance of joint audio-visual modeling. We would also like to underline that—while some promising results from bi-modal words perspective have been shown in this paper—advanced joint audio-visual representations is still a topic that deserves more in-depth studies in the future. It is also interesting and important to construct a larger dataset for evaluating these new representations.

ACKNOWLEDGMENT

This work is supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via 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. Disclaimer: 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. Yu-Gang Jiang is supported by two grants from NSF China (#61201387 and #61228205), two grants from STCSM (#12XD1400900 and #12511501602), and a New Teachers Fund for Doctoral Stations, Ministry of Education (#20120071120026), China.

References

1. http://en.wikipedia.org/wiki/P-value.
2. http://www.nist.gov/itl/iad/mig/med10.cfm/.
3. http://www.nist.gov/itl/iad/mig/med11.cfm/.
4. L. Bao, et al. Informedia @ TRECVID 2011. In *TRECVID Workshop*, 2011.
5. M. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Transcations on Pattern Analysis and Machine Intelligence*, 2003.
6. Y.-L. Boureau, J. Ponce, and Y. Lecun. A theoretical analysis of feature pooling in visual recognition. In *International Conference on Machine Learning*, 2010.
7. G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *European Conference on Computer Vision*, 2004.

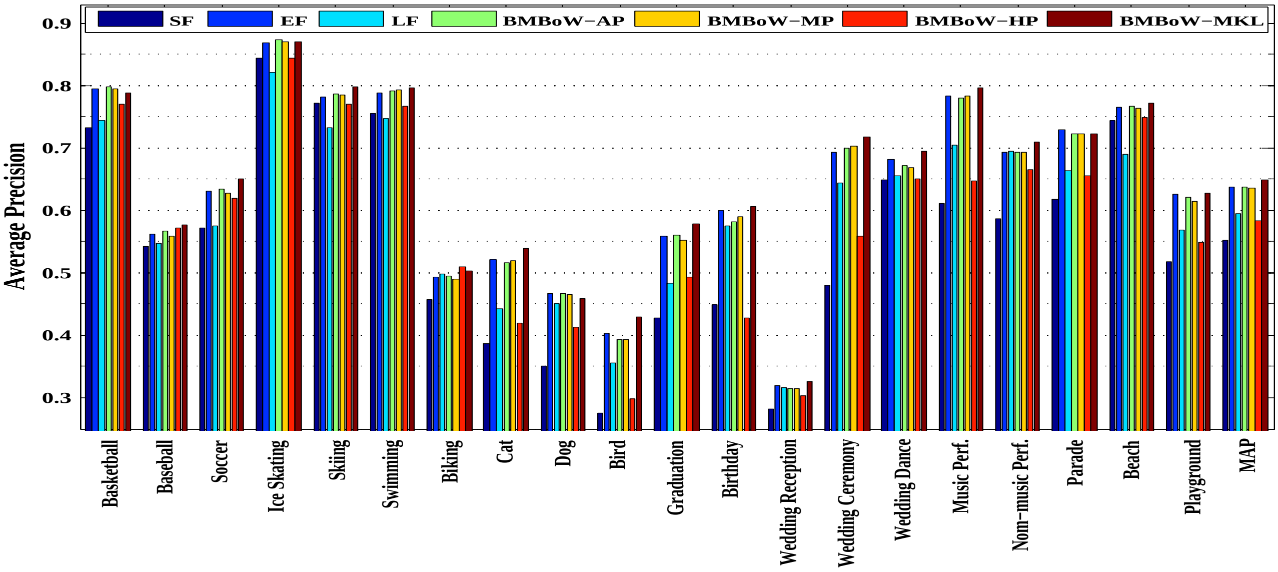


Fig. 9 Per-category performance comparison on CCV dataset. This figure is best viewed in color.

8. M. Cristani, M. Bicego, and V. Murino. Audio-visual event recognition in surveillance video sequences. *IEEE Transactions on Multimedia*, 2007.
9. I. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *ACM Conference on Knowledge Discovery and Data Mining*, 2001.
10. P. Gehler and S. Nowozin. On Feature Combination for Multiclass Object Detection. In *IEEE International Conference on Computer Vision*, 2009.
11. I.H. Jhuo and D.-T. Lee. Boosting-based Multiple Kernel Learning for Image Re-ranking. In *ACM International Conference on Multimedia*, 2010.
12. W. Jiang, C. Cotton, S.-F. Chang, D. Ellis and A. Loui. Short-term audio-visual atoms for generic video concept classification. In *ACM International Conference on Multimedia*, 2009.
13. W. Jiang and A. Loui. Audio-visual grouplet: Temporal audio-visual interactions for general video concept classification. In *ACM International Conference on Multimedia*, 2011.
14. Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis and A. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ACM International Conference on Multimedia Retrieval*, 2011.
15. Y.-G. Jiang, et al. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010.
16. Y.-G. Jiang, S. Bhattacharya, S.-F. Chang and M. Shah High-Level Event Recognition in Unconstrained Videos. In *International Journal of Multimedia Information Retrieval*, 2012.
17. A. Kembhavi, B. Siddiquie, R. Miezianko, S. McCloskey and Larry S. Davis. Incremental Multiple Kernel Learning for Object Recognition. In *IEEE International Conference on Computer Vision*, 2009.
18. D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.
19. I. Laptev and T. Lindeberg. On space-time interest points. *International Journal of Computer Vision*, 2005.
20. I. Laptev, M. Marszlek, C. Schmid and B. Rozenfeld. Learning realistic human actions from movies. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.
21. J. Liu, M. Shah, B. Kuipers, and S. Savarese. Cross-view action recognition via view knowledge transfer. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.
22. H. Lutkepohl. Handbook of Matrices. In *Chichester: Wiley*, 1997.
23. C. Manning, P. Raghavan, and H. Schtze. Introduction to information retrieval. *Cambridge University Press*, 2008.
24. K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 2004.
25. P. Natarajan et al. BBN VISER TRECVID 2011 Multimedia Event Detection System. In *NIST TRECVID Workshop*, 2011.
26. S. Pan, X. Nu, J. T. Sun, Q. Yang, and Z. Chen. Co-clustering documents and words using bipartite spectral graph partitioning. In *International World Wide Web Conference*, 2010.
27. L. Pols. Spectral analysis and identification of Dutch vowels in monosyllabic words. *Doctoral dissertion, Free University, Amsterdam*, 1966.
28. G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: an overview. In *Issues in visual and audio-visual speech processing*, 2004.
29. A. Rakotomamonjy, F.R. Bach, S. Canu and Y. Grandvalet. SimpleMKL. In *Journal of Machine Learning Research*, 2009.
30. A. Vedaldi, V. Gulshan, M. Varma and A. Zisserman. Multiple kernels for object detection. In *IEEE International Conference on Computer Vision*, 2009.
31. J.-C. Wang, Y.-H. Yang, I.-H. Jhuo, Y.-Y. Lin and H.-M. Wang The Acousticvisual Emotion Guassians Model for Automatic Generation of Music Video. In *ACM International Conference on Multimedia*, 2012.
32. G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang. Robust late fusion with rank minimization. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.
33. G. Ye, I.-H. Jhuo, D. Liu, Y.G. Jiang, D.-T. Lee and S.-F. Chang. Joint Audio-Visual Bi-Modal Codewords for Video Event Detection. In *ACM International Conference on Multimedia Retrieval*, 2012.