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

Guangnan Ye[†], I-Hong Jhuo[‡], Dong Liu[†], Yu-Gang Jiang[§], D. T. Lee[‡♯], Shih-Fu Chang[†]
[†]Dept.of Electrical Engineering, Columbia University
[‡]Dept.of Computer Science and Information Engineering, National Taiwan University
[§]School of Computer Science, Fudan University
[♯]Dept. of Computer Science and Engineering, National Chung Hsing University
{yegn,dongliu,sfchang}@ee.columbia.edu, ihjhuo@gmail.com, ygj@fudan.edu.cn, dtlee@ieee.org

ABSTRACT

Joint audio-visual patterns often exist in videos and provide strong multi-modal cues for detecting multimedia events. However, conventional methods generally fuse the visual and audio information only at a superficial level, without adequately exploring deep intrinsic joint patterns. In this paper, we propose a joint audio-visual bi-modal representation, called bi-modal words. 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 construct the bi-modal words that reveal the joint patterns across modalities. Finally, different pooling strategies are employed to re-quantize the visual and audio words into the bi-modal words and form bi-modal Bag-of-Words representations that are fed to subsequent multimedia event classifiers. We experimentally show that the proposed multi-modal feature achieves statistically significant performance gains over methods using individual visual and audio features alone and alternative multi-modal fusion methods. Moreover, we found that average pooling is the most suitable strategy for bi-modal feature generation.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithm, Experimentation

Keywords

Bi-Modal Words, Event Detection, Feature Pooling

1. INTRODUCTION

Automatic detection of complex multimedia events in Internet videos has great potential for many applications, such as general video retrieval, consumer content management, etc. A large portion of the Internet videos are captured and uploaded by ordinary consumers. These videos are typically recorded under uncontrolled conditions without professional post-editing, showing large variations in lighting, viewpoint and camera motion. Figure 1 gives several example frames of videos containing the same event.

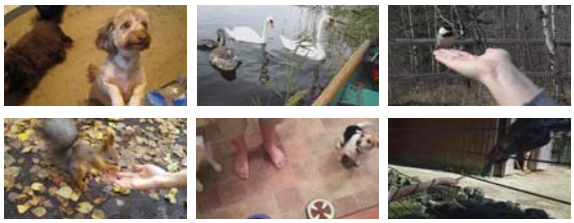

Figure 1: Event detection in unconstrained videos is a challenging task due to uncontrolled conditions such as lighting, occlusions, and complicated camera motions, etc. The content is extremely diverse as shown in the image frames extracted from the "feeding an animal" event category defined in TRECVID MED 2011.

Notably, events captured in the videos are implicitly multi-modal 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., runner with cheering sounds in baseball videos). Therefore, we believe that successful event detection solutions should harness both audio and visual modalities. Also, we expect the cross-modal correlation to be best preserved in consumer videos due to the raw audio-visual content seen in such videos. Such correlations might disappear after editing seen in other domains such as broadcast, commercials, or videos used in social media sites.

The existing audio-visual video event analysis approaches evolve through three paradigms. The first is early fusion [3], which combines the audio and visual information before performing classification, such as feature concatenation or multiple kernel learning. The second paradigm is late fusion [12, 22], which aims at combining the prediction scores of the individual models constructed from audio or vi-

ICMR '12, June 5-8, Hong Kong, China

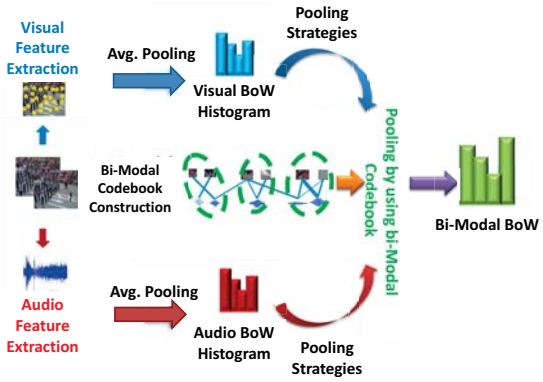

Figure 2: The generation process of the proposed audio-visual bi-modal BoW representation. First, audio and visual features are extracted from the videos and quantized into audio and visual BoW histogram respectively. Then a bipartite graph is constructed to model the relations across the quantized words extracted from the visual and audio modalities, in which each node denotes a visual or audio word and each edge between two nodes encodes the correlation between the two words. By partitioning the bipartite graph into a number of clusters, we obtain several bi-modal words that reveal the joint audio-visual patterns. Finally, the audio and visual words in the original BoW representations are re-quantized into the bi-modal words with different pooling strategies.

sual information. Both paradigms combine the information from different modalities only in a shallow manner, lacking through exploitation of their deep correlations. In the third paradigm, deep multi-modal analysis models are proposed to discover the joint audio-visual patterns in the videos [9, 10]. Nevertheless, such models typically need to perform object or region tracking followed by complex audio-visual joint modeling, which prevents their applicability in the real world unconstrained videos with complex object and scene interactions.

We propose an audio-visual bi-modal Bag-of-Words (BoW) representation that describes the video contents based on the joint relations between the audio and visual modalities. The framework is illustrated in Figure 2. First, we apply BoW approach to build audio words and visual words through the standard k-means clustering method separately. Then, a bipartite graph is constructed to capture joint statistics between the quantized audio words and visual words. After that, the spectral clustering method is used for bipartite graph partitioning. Finally, the original individual words in each modality (audio, or visual) are re-quantized into a number of bi-modal codewords, which are employed as the audio-visual bi-modal BoW representation. In addition, we evaluate different pooling strategies in the re-quantization stage, and show that average pooling is an effective strategy in the audio-visual bi-modal BoW construction.

The audio-visual bi-modal BoW offers several distinct advantages: (1) Easy to implement, in which only a bipartite graph partitioning procedure is needed to obtain the bi-modal words. Moreover, the graph-based representation is a good choice for uncovering the manifold structure underlying the joint audio-visual feature space, and thus provides unique representations distinct from those in the prior work; (2) The dimensionality of the features can be greatly reduced (e.g. from about $14,000$ to $4,000$), and thereby can result in a better generalization; (3) It also provides strong multi-modal cues and discriminative power for detecting the multimedia events. From the experiment results shown in Section 5, our method is able to achieve the best performance over the state of the arts in the task of video event detection.

2. RELATED WORK

The most popular audio-visual analysis strategy is multi-modality fusion, which tries to fuse the audio and visual information in a complementary manner. For example, the work in [3] averaged the kernel matrices obtained from the audio and visual features as the early fusion method. Jiang *et al.* [12] trained independent event classifiers based on BoW representation of the audio or visual feature, and then performed late fusion to combine the prediction results of the classifiers obtained from different modalities. Different from such fusion methods, we pursue a novel bi-modal feature representation that characterizes the joint patterns across the audio and visual modalities, which better captures the underlying relations between low-level features and high-level semantics.

Related work can also be found in audio-visual speech recognition [21]. But they are restricted to videos of talking faces and only special features associated with faces and speeches are used. Note our work focuses on general situations without the limitation to talking faces and speech recognition only.

There are also some work that explore the joint audio-visual analysis for object detection and tracking in the literature. For example, Cristani *et al.* [7] proposed to synchronize foreground of visual objects and the background of audio sounds for object detection. Beal *et al.* [4] proposed a joint probability model of both audio and visual information for tracking the object motion. Nevertheless, these methods are only designed for videos in a controlled environment in which the foreground and background can be easily separated. However, such requirement cannot be satisfied in detecting events from uncontrolled videos with complex contextual interactions among different semantic entities. Recently, Jiang *et al.* [9] developed a Short-Time Audio-Visual Atom (ST-AVA) as the joint audio-visual feature for video concept classification. First, visual regions are tracked within the short-term video slices to generate the 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 the AVA feature representation. Finally, a discriminative joint audio-visual codebook is constructed from the AVAs using multiple instance learning, and the codebook-based features are generated for semantic concept detection. As an extension of this work, Jiang *et al.* [10] further proposed the Audio-Visual Grouplet (AVG) by exploring the temporal audio-visual interactions, in which an AVG 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 AVGs 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 works need to perform object or region tracking. However, since the events typically consists of complex interactions between objects and scenes, it will be difficult, if not impossible, to perform effective and efficient object tracking in unconstrained videos. In addition, tracking also incurs significant computational burden in practice. Therefore, such tracking based methods may not be adequate for capturing the audio-visual patterns in the events, which directly motivates our work in this paper.

Methodologically, our work is motivated by the bipartite graph partitioning technique [8] which has been widely applied in various applications. For example, Pan *et al.* [19] constructed a bipartite graph to model the co-occurrence relations between words in different domains, and then adapted 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. Liu *et al.* [15] used the bipartite graph to model the co-occurrence relation of two view-dependent visual vocabularies, and applied the graph partitioning to find visual word co-clusters. The co-clusters can transfer view knowledge across different views, and realize the cross-view action recognition. In contrast to these applications which focus on cross-domain/view learning, we use the bipartite graph to discover the correlations between audio and visual words, which greatly reduces the dimensionality of the features and provides strong cues and discriminative power for detecting events.

3. FEATURE EXTRACTION FROM INDIVIDUAL MODALITIES

We follow the BoW feature representations in [12] which typically forms the feature representations by the following four steps. First, a set of key points are sampled from the videos by detecting the 2D/3D local features [13, 14] in the videos. Second, we extract a descriptor from each detected key point that represents the local appearance/motion at different locations in the videos. Next, these descriptors are quantized into a codebook of feature vectors, typically with k-means clustering method. Finally, the quantizations are aggregated to form a single, fixed-dimensional histogram to represent the video. In this work, we extract three kinds of low-level features from the video contents, which are described as follows:

SIFT Appearance Feature. SIFT has been proved very effective especially for object and scene categorization. Following [12], we adopt two versions of sparse keypoint detector: Different of Gaussian [13], and Hessian Affine [17], to find local keypoints. Each keypoint is described by a 128 dimensional vector. To reduce the computational processing time, we sample one frame from every two seconds of video. The SIFT features within a frame are further quantized into a $5,000$-dimensional BoW histogram.

STIP Motion Feature. Unlike SIFT, STIP captures motion information in the video. It extracts space-time local volumes which have significant variations in space and time. We apply Laptev's method [14] to compute the keypoints and the corresponding descriptors of STIP. The descriptors are computed from HOG and HOF (144 dimensions). The STIP features within the local volume are quantized into a $5,000$-dimensional BoW histogram.

MFCC Audio Feature. Except for the visual features such as SIFT and STIP, audio information is also very important for event classification. We extract the popular Mel-frequency cepstral coefficients (MFCC) [20] as the audio feature in this work. We compute the 60-dimensional MFCC feature for every window of 32ms with 50% overlap, and then apply k-means method on these MFCC features to form the codebook of size $4,000$. Finally, the MFCC feature within the time window is further quantized into a $4,000$-dimensional BoW histogram.

In the experiments of this work, we combine the visual words generated from SIFT and STIP together as the visual codebook ($5,000$+$5,000 = 10,000$ codewords) while treating the $4,000$ audio words as the audio codebook. It is worth noting that the above three BoW representations are local features which represent the keyframe or local volume of the video clip. To obtain an aggregated single feature vector for the entire video clip, we use average pooling to obtain the video level feature representation, i.e., taking the average of the BoW features of all the keyframes or local volumes.

4. AUDIO-VISUAL BI-MODAL REPRESENTATION

In this section, we present our audio-visual bi-modal representation for video event detection. We first describe the bipartite graph construction for modeling the correlation of audio and visual words, and then introduce how to generate the bi-modal words through bipartite graph partitioning. Finally, we discuss several pooling strategies for re-quantizing the original visual/audio BoW into the audio-visual bi-modal BoW representation.

4.1 Bipartite Graph Construction

Assume that we are given a training video collection $\mathcal{D} = \{d_i\}_{i=1}^n$ with n videos. Each video d_i is represented as a bi-modal representation $d_i = \{\mathbf{h}_i^a, \mathbf{h}_i^v\}$, where $\mathbf{h}_i^a$ denotes the $4,000$-dimensional audio BoW feature and $\mathbf{h}_i^v$ denotes the $10,000$-dimensional visual BoW feature. We use ℓ_1 normalization on each of the above BoW feature representations such that the sum of its entries equals 1. For simplicity, we use $\mathcal{W}^a = \{w_1^a, \ldots, w_{m_a}^a\}$ and $\mathcal{W}^v = \{w_1^v, \ldots, w_{m_v}^v\}$ to denote the sets of audio and visual words respectively, where $w_i^a \in \mathcal{W}^a$ and $w_i^v \in \mathcal{W}^v$ represent one audio word and one visual word, m_a and m_v denotes the number of audio and visual words. The total number of audio and visual words can be denoted as $m = m_a + m_v$.

Based on the given training video collection, we can construct a bipartite 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. Specifically, the vertex set V is a finite set $V = V^a \bigcup V^v$, 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 edges connecting two vertices in V^a or V^v respectively. For any edge $e_{kl} \in E$, we associate a non-negative weight s_{kl} to measure the correlation between audio word $w_k^a \in \mathcal{W}^a$ and

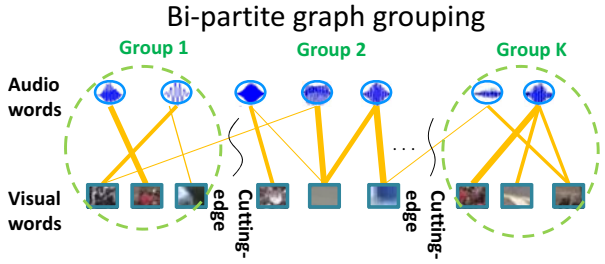

Figure 3: An illustration of bipartite graph constructed between audio and visual words, where the upper vertices denote the audio words and the lower vertices denote the visual words. Each edge connects one audio word and one visual word, which is weighted by the correlation measure calculated based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

visual word $w_l^v \in \mathcal{W}^v$, 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 audio word w_k^a and $\mathbf{h}_i^v(l)$ denotes the entry of $\mathbf{h}_i^v$ corresponding to video word w_l^v.

Now we explain the rationality of Eq. (1). Specifically, the numerator measures the summation of the joint probability of audio word w_k^a and visual word w_l^v, where the summation is calculated over the entire video collection. This value essentially reveals the correlation of audio and visual words, in which the higher the value, the more correlative for these two words. On the other hand, the denominator acts as a normalization term, which penalizes the audio and/or visual words that frequently appeared in the video collection. It is worth noting that the underlying principle in Eq. (1) is similar to the term of *tf-idf* in information retrieval, which has proven to be effective in various applications [16]. Figure 3 illustrates a bipartite graph constructed from the joint statistic of the audio and visual words.

Note that the choice of correlation measure in Eq. (1) is flexible and we can also adopt other methods to estimate s_{kl}, e,g., the Pointwise Mutual Information (PMI) in [15]. In addition, the co-occurrence measure (the numerator in Eq. (1)) can be computed over shorter temporal durations by segmenting each video into shorter clips.

4.2 Bi-Modal Words Discovery

Based on the bipartite graph constructed between audio and visual words, we adopt the bipartite graph partitioning method to discover the audio-visual bi-modal words. Following the discussion in [8], we begin with bipartitioning method over the bipartite graph and then extend it into the multipartitioning scenario.

Suppose 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 can be defined as sum of all edge weights connecting vertices in two subsets,

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

The bipartitioning 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,

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

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

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

Based on the above definitions, it can be shown that the graph cut can be equally written as the following equivalent 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 the 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 $\mathbf{L}\mathbf{z} = \lambda \mathbf{D}\mathbf{z}$ (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). To obtain the the eigenvector corresponding to the second smallest eigenvalue, [8] 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

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.

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 $\mathcal{C}_j$ $(j = 1, 2)$.

In practice, the above objective function can be 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 [8]. Suppose we have $l = \lceil \log_2 K \rceil$ singular vectors $\mathbf{u}_2, \mathbf{u}_3, \ldots, \mathbf{u}_{l+1}$, and $\mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_{l+1}$, 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., audio or visual words corresponding to the background).

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 on the MATLAB platform of a Six-Core Intel Xeon Processor X5660 with 2.8 GHz CPU and 32 GB memory, and observe that it takes 32 minutes to cluster $14,000$ audio and visual words into $2,000$ bi-modal words in the experiment on CCV dataset (see Section 5.2).

4.3 Audio-Visual BoW Generation

After obtaining the bi-modal words, we need to re-quantize the original audio and visual BoW representations into the bi-modal words such that the videos can be represented as the audio-visual BoW representation. For any video $d_i = (\mathbf{h}_i^a, \mathbf{h}_i^v)$, we consider three quantization strategies for generating the audio-visual BoW representation, which are described as follows respectively:

Average Pooling. This audio-visual bi-modal BoW generation strategy is formally 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 $\mathbf{h}_i^{\text{avg}}(k)$ 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. From Eq.(12), we can see that the entry of the bi-modal representation is the average value of the entries corresponding to 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 [5].

Max Pooling. The second strategy is the max pooling, which is formally defined as follows,

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

where essentially selects the largest summation in the original audio or visual words as the quantization value of the given audio-visual bi-modal word.

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 tends 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 shows the generation procedure of the audio-visual BoW representation. Once we obtain the audio-visual BoW representations of the video collection, we can use these features and the label information on the training set to train event classifiers. Finally, the event detection is performed by applying the learnt classifiers on the test videos.

5. EXPERIMENTS

In this section, we evaluate our proposed audio-visual bi-modal representation on various benchmark datasets for video event detection. As discussed in Section 3, we apply 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. The following audio/visual feature representations will be compared: (1) Single Feature (SF), where we only report the best performance achieved by one of the three features mentioned above. (2) Early Fusion (EF). We concatenate three kinds of BoW features into a long vector

with the dimensions of 14,000. (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 scores for event detection. (4) Average Pooling based Bi-Modal BoW (BMBoW-AP), where the average pooling strategy is employed to generate the bi-modal BoW. (5) Max Pooling based Bi-Modal BoW (BMBoW-MP), in which we use max pooling to generate the audio-visual BoW. (6) Hybrid Pooling based Bi-Modal BoW (BMBoW-HP), which applies the hybrid pooling described in Section 4.3.

We use the one-vs-all SVM as the classifier for event detection, in which the positive videos labeled with the given event and the other negative videos that do not belong to the concerned event are used as training data for training the SVM classifier of the target event. We employ the Average Precision (AP) as the evaluation metric of event detection. We calculate AP for each event and then calculate the Mean Average Precision (MAP) across all the events of the dataset as the final evaluation metric.

For the parameter setting of the SVM classifier, we vary the tradeoff parameter C of SVM on the grid of $\{10^{-1} \dots 10^{3}\}$ and then choose the best value based on validation performance. To get the optimal parameter for each method, we partition the training set into 10 subsets and then perform 10-fold cross validation. Moreover, we apply χ^2 kernel as the kernel matrix for SVM classifier, which is calculated as $k(x,y) = e^{-\frac{d_{\chi^2}(x,y)}{\sigma}}$ where σ is by default set as the mean value of all pairwise distances on the training set.

5.1 Experiment on TRECVID MED 2011 Development Dataset

TRECVID MED is a challenging task for the detection of complicated high-level events. We test our proposed method on TRECVID MED 2011 development dataset [2], which includes five events "Attempting a board trick", "Feeding an animal", "Landing a fish", "Wedding ceremony", and "Working on a woodworking project" and one background class. This dataset consists of 10,804 videos from 17,566 minutes of web videos, which is partitioned into the training set (8,783 videos) and the test set (2,021 videos).

Figure 4 shows the per-event performance for all the methods in comparison, where the bi-modal codebook size is set as 4,000. From the results, we have the following observations: (1) Our proposed audio-visual bi-modal BoW representation produces better result than all the other baseline methods in terms of MAP, with significant performance improvements on all of the five events. (2) The audio-visual BoW representation outperforms the early fusion and late fusion methods by a large margin. This is due to the fact that the bi-modal words capture the correlation between audio and visual information while the latter two methods only fuse audio and visual information in an aggregated manner without exploring their mutual dependence. (3) The bi-modal feature performs significantly better than all the single feature, which verifies the merits of considering multi-modality in the task of video event detection. (4) BMBoW-AP tends to produce 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 the suitable pooling strategies for different modalities (i.e., max pooling for the transient audio signal and average pooling for the persistent visual signal). In addition, BMBoW-HP even achieves better results than BMBoW-AP (see Fig. 5 in the following paragraph), especially when the size of bi-modal code words is relative large, indicating that selecting maximum response of audio signal may help reveal the semantic clue of the videos.

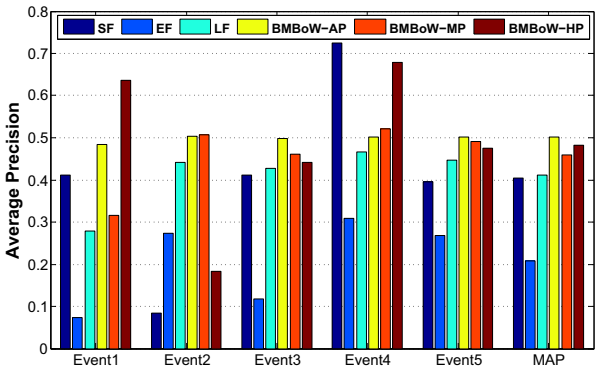

Figure 4: Per-event performance on TRECVID MED 2011 video event detection. The five event from left to right in the horizontal axis are "Attempting board trick", "Feeding an animal", "Landing a fish", "Wedding ceremony", and "Wood working", and the final result is the MAP. This figure is best viewed in color.

Figure 5 further shows the MAP performance of different pooling strategies when the size of the audio-visual bi-modal codebook varies from 2,000 to 12,000. As seen, average pooling tends to enjoy 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. Figure 6 shows the density of audio and visual words within each bi-modal word. Each point in the map denotes the frequency of bi-modal words made up of a certain numbers of audio word (vertical coordinate, only up to 18 is shown in the figure) and visual word (horizontal coordinate, only up to 22 is shown in the figure). It estimates the portion of the words in the entire bi-modal codebook that contain both visual and audio information, which was found to be about 47% for the TRECVID MED 2011 data set. This confirms the significant effect of the bi-modal correlations in the joint multi-modal representation. The bi-modal feature was also an important component of the large feature set used in our system and achieved the best performance [18] in TRECVID MED 2011.

5.2 Experiment on Columbia Consumer Video (CCV) Dataset

In the second experiment, we use the Columbia Consumer Video (CCV) dataset [11], This dataset 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. To facilitate benchmark comparison, we report performance of all the 20 categories.

Figure 7 shows the per-category performance comparison of all the methods, where the bi-modal codebook size is set as 6,000. From the results, we can see that the proposed BMBoW-AP achieves the best performance in terms

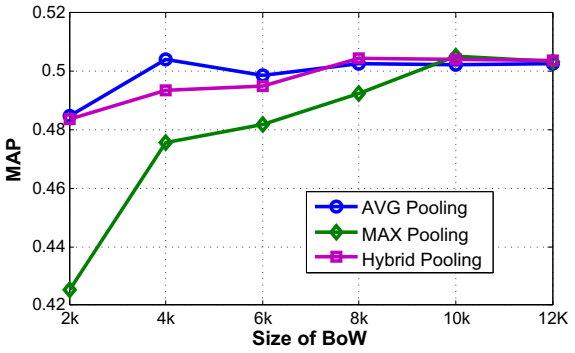

Figure 5: Effect of varying bi-modal codebook size on TRECVID MED 2011 performance.

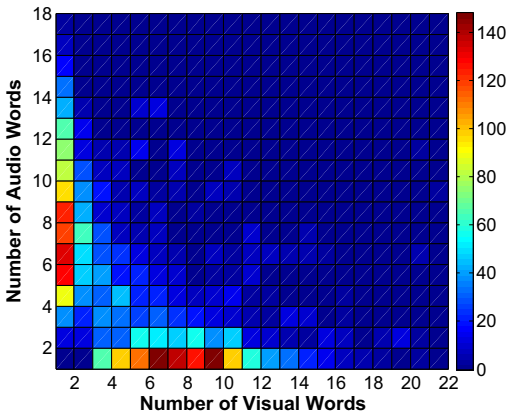

Figure 6: The density of audio and visual words within the bi-modal words on TRECVID MED 2011 development dataset, where the total number of bi-modal words is set as $4,000$. Each point in the map denotes the frequency of bi-modal words made up of a certain numbers of audio word (vertical coordinate, only up to 18 is shown in the figure) and visual word (horizontal coordinate, only up to 22 is shown in the figure).

of MAP. In particular, it outperforms the BMBoW-MP, BMBoW-HP and LF by 1.1%, 6.2% and 5.1% respectively, which clearly demonstrates that our method is superior to all the baseline feature representations. Moreover, it achieves the best performance on most of the event categories. For instance, on event "graduation", our method outperforms the best baseline method SF by 14.2%. Besides, comparing with the the best baseline EF, our method achieves the highest relative performance gain on category "bird " and "wedding ceremony". The reason may be that these two categories contain more significant audio-visual correlation than the other categories. For example, the appearance of birds are often accompanied with the birds' singing audio background. Meanwhile, people's actions in wedding ceremony are always with background music. In general, we expect high impact of the proposed bi-modal features on other evens that share strong audio-visual correlations like the ones mentioned above.

Figure 8 further shows MAP performance of different pooling strategies with variant sizes of the codebook. As can be seen, the average pooling achieves significant and consistent MAP results as the codebook size varies. Figure 9 shows the density of audio and visual word within each bi-modal word for the CCV dataset, for which about 36% of bi-modal codewords contain both audio and visual codewords.

We also measure the statistical significance between the best baseline and BMBoW-AP on the two datasets. A popular measure for statistical significance testing, the p-value, 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 $D_i = MAP_{\text{BMBoW-AP}}(i) - MAP_{\text{Baseline}}(i)$, 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 and CCV dataset are 0.019 and 0.022 respectively, which are well below 0.05 and shows that the null hypothesis can be rejected. Therefore, we conclude that our method has achieved statistical significant improvements over the best baseline on the two datasets.

6. CONCLUSION

We have introduced an audio-visual bi-modal representation for video event detection. The proposed method discovers the joint audio-visual patterns in the videos by the bipartite graph partitioning. Different pooling strategies are employed to re-quantize the audio and visual BoW representations into the bi-modal words, where average pooling is found to be most suitable for bi-modal BoW generation. Extensive experiments have demonstrated the effectiveness of the proposed method on video event detection. For future work, we will consider the following directions: (1) The bi-modal representations provide a common framework for measuring the "similarity" between features of different modalities, and thus can be used for cross-modal retrieval. (2) We will study the class-dependent bi-modal words to explore the audio-visual patterns that are unique to each individual event.

7. ACKNOWLEDGMENT

This work is supported 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.

8. REFERENCES

[1] http://en.wikipedia.org/wiki/P-value.
[2] http://www.nist.gov/itl/iad/mig/med11.cfm/.

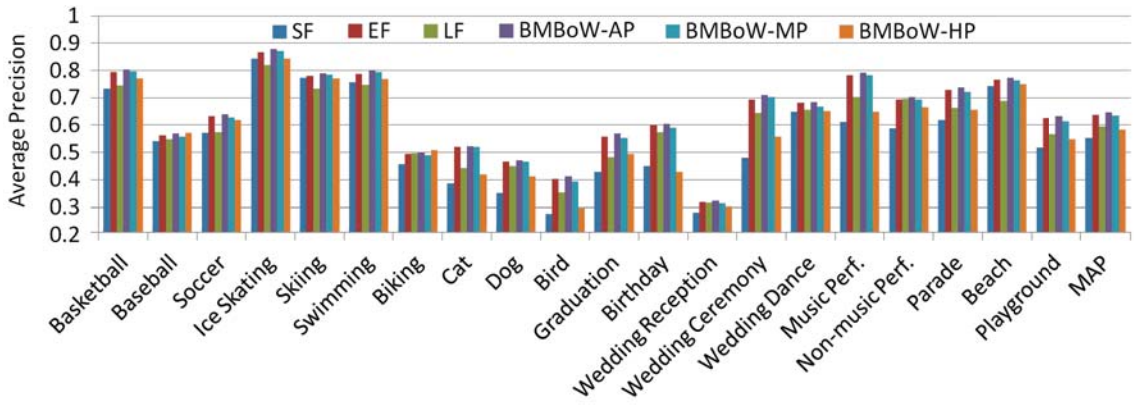

Figure 7: Per-category performance comparison on CCV dataset. This figure is best viewed in color.

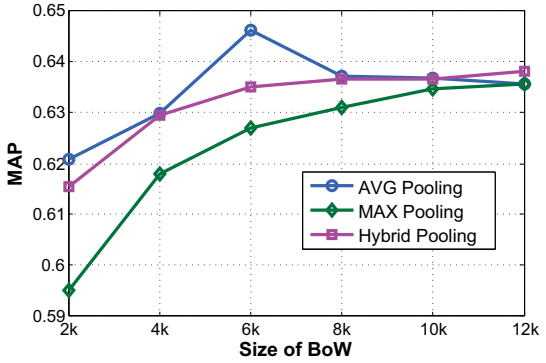

Figure 8: Effect of varying bi-modal codebook size on CCV dataset performance.

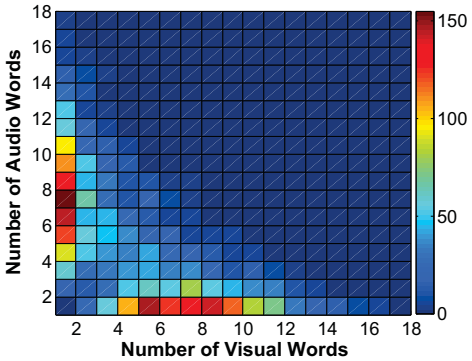

Figure 9: The density of audio and visual words within the bi-modal word on CCV dataset, where the number of bi-modal words is $6,000$ (vertical coordinate up to 18 is shown, horizontal coordinate up to 22 is shown in the figure).

[3] L. Bao, et al. Informedia @ TRECVID 2011. In *TRECVID Workshop*, 2011.

[4] M. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *TPAMI*, 2003.

[5] Y.-L. Boureau, J. Ponce, and Y. Lecun. A theoretical analysis of feature pooling in visual recognition. In *ICML*, 2010.

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV*, 2004.

[7] M. Cristani, M. Bicego, and V. Murino. Audio-visual event recognition in surveillance video sequences. *TMM*, 2007.

[8] I. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *SIGKDD*, 2001.

[9] W. Jiang, C. Cotton, S.-F. Chang, D. Ellis and A. Loui. Short-term audio-visual atoms for generic video concept classification. In *MM*, 2009.

[10] W. Jiang and A. Loui. Audio-visual grouplet: Temporal audio-visual interactions for general video concept classification. In *MM*, 2011.

[11] 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 *ICMR*, 2011.

[12] Y.-G. Jiang, et al. Columbia-ucf trecvid2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. In *NIST TRECVID Workshop*, 2010.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[14] I. Laptev and T. Lindeberg. On space-time interest points. *IJCV*, 2005.

[15] J. Liu, M. Shah, B. Kuipers, and S. Savarese. Cross-view action recognition via view knowledge transfer. In *CVPR*, 2011.

[16] C. Manning, P. Raghavan, and H. Schtze. Introduction to information retrieval. *Cambridge University Press*, 2008.

[17] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 2004.

[18] P. Natarajan et al. BBN VISER TRECVID 2011 Multimedia Event Detection System. In *In NIST TRECVID Workshop*, 2011.

[19] S. Pan, X. Nu, J. T. Sun, Q. Yang, and Z. Chen. Co-clustering documents and words using bipartite spectral graph partitioning. In *WWW*, 2010.

[20] L. Pols. Spectral analysis and identification of Dutch vowels in monosyllabic words. *Doctoral dissertion, Free University, Amsterdam*, 1966.

[21] 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.

[22] G. Ye, D. Liu, I.-H. Jhuo, and S.-F. Chang. Robust late fusion with rank minimization. In *CVPR*, 2012.