

# Audio-Visual Grouplet: Temporal Audio-Visual Interactions for General Video Concept Classification

Wei Jiang Alexander C. Loui

Corporate Research and Engineering, Eastman Kodak Company, Rochester, NY  
{wei.jiang, alexander.loui}@kodak.com

## ABSTRACT

We investigate general concept classification in unconstrained videos by joint audio-visual analysis. A novel representation, the Audio-Visual Grouplet (AVG), is extracted by studying the statistical temporal audio-visual interactions. An AVG is defined as a set of audio and visual codewords that are grouped together according to their strong temporal correlations in videos. The AVGs carry unique audio-visual cues to represent the video content, based on which an audio-visual dictionary can be constructed for concept classification. By using the entire AVGs as building elements, the audio-visual dictionary is much more robust than traditional vocabularies that use discrete audio or visual codewords. Specifically, we conduct coarse-level foreground/background separation in both audio and visual channels, and discover four types of AVGs by exploring mixed-and-matched temporal audio-visual correlations among the following factors: visual foreground, visual background, audio foreground, and audio background. All of these types of AVGs provide discriminative audio-visual patterns for classifying various semantic concepts. We extensively evaluate our method over the large-scale Columbia Consumer Video set. Experiments demonstrate that the AVG-based dictionaries can achieve consistent and significant performance improvements compared with other state-of-the-art approaches.

## Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing; H.4.m [Information Systems Applications]: Miscellaneous; H.3.m [Information Storage and Retrieval]: Miscellaneous

## General Terms

Algorithms, Experimentation

## Keywords

Audio-visual grouplet, video concept classification, temporal audio-visual interaction, foreground/background separation

\*Area chair: Cees G.M. Snoek

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

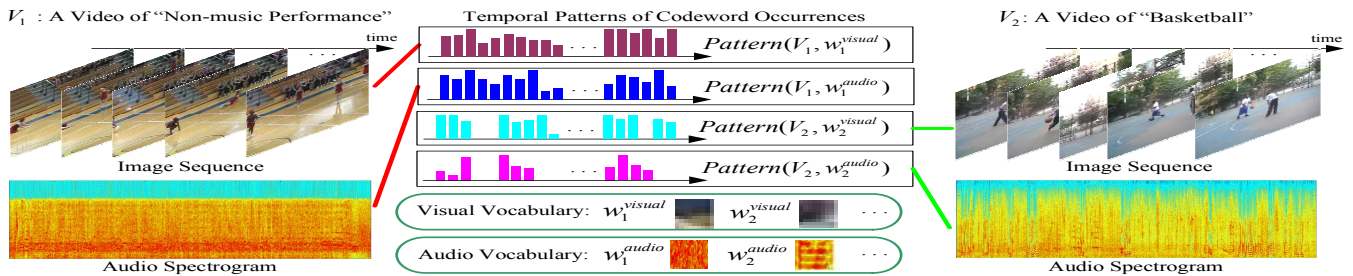
Copyright 2011 ACM 978-1-4503-0616-4/11/11 ...\$10.00.

## 1. INTRODUCTION

This paper investigates the problem of automatic classification of semantic concepts in generic, unconstrained videos, by joint analysis of audio and visual content. These concepts include general categories, such as scene (*e.g.*, beach), event (*e.g.*, birthday, graduation), location (*e.g.*, playground) and object (*e.g.*, dog, bird). Generic videos are captured in an unrestricted manner, like those videos taken by consumers on YouTube. This is a difficult problem due to the diverse video content as well as the challenging conditions such as uneven lighting, clutter, occlusions, and complicated motions of both objects and camera.

Large efforts have been devoted to classify general concepts in generic videos, such as TRECVID high-level feature extraction [32] and Columbia Consumer Video (CCV) concept classification [18]. Most previous works classify videos in the same way they classify images, using mainly visual information. Specifically, visual features are extracted from either 2D keyframes or 3D local volumes, and these features are treated as individual static descriptors to train concept classifiers. Among these methods, the ones using the Bag-of-Words (BoW) representation over 2D or 3D local descriptors (*e.g.*, SIFT [23] or HOG [8]) are considered state-of-the-art, due to the effectiveness of BoW features in classifying objects [13], events [32], and human actions [21].

The importance of incorporating audio information to facilitate semantic concept classification has been discovered by several previous works [5, 18, 37]. They generally use a multi-modal fusion strategy, *i.e.*, early fusion [5, 18, 37] to train classifiers with concatenated audio and visual features, or late fusion [5, 37] to combine judgments from classifiers built over individual modalities. Different from such fusion approaches that avoid studying temporal audio-visual synchrony, the work in [16] pursues a coarse-level audio-visual synchronization through learning a joint audio-visual codebook based on atomic representations in both audio and visual channels. However, the temporal audio-visual interaction is not explored in previous video concept classification methods. The temporal audio-visual dependencies can reveal unique audio-visual patterns to assist concept classification. For example, as illustrated in Figure 1, by studying the correlations between temporal patterns of visual and audio codewords, we can discover discriminative audio-visual cues, such as the unique encapsulation of visual basketball patches and audio basketball bouncing sounds for classifying the “basketball” concept, and the encapsulation of visual stadium patches and audio music sounds for classifying



**Figure 1: Discovery of audio-visual patterns through temporal audio-visual interactions.**  $w_i^{visual}$  and  $w_j^{audio}$  are discrete codewords in visual and audio vocabularies, respectively. By analyzing the correlations between the temporal histograms of audio and visual codewords, we can discover salient audio-visual cues to represent videos from different concepts. For example, the highly correlated visual basketball patches and audio basketball bouncing sounds provide a unique pattern to classify “basketball” videos. The highly correlated visual stadium patches and audio background music are helpful to classify “non-music performance” videos. In comparison, discrete audio and visual codewords are less discriminative than such audio-visual cues.

the “non-music performance” concept<sup>1</sup>. To the best of our knowledge, such audio-visual cues have not been studied before in previous literatures.

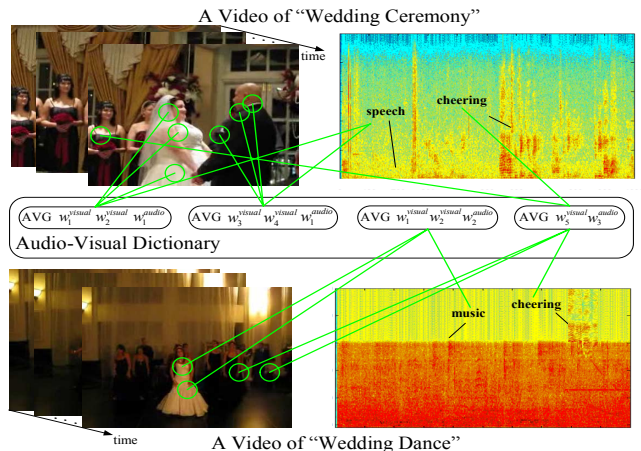
From another perspective, beyond the traditional BoW representation, structured visual features have been recently found to be effective in many computer vision tasks. In addition to the local feature appearance, spatial relations among the local patches are incorporated to increase the robustness of the visual representation. The rationale behind this is that individual local visual patterns tend to be sensitive to variations such as changes of illumination, views, scales, and occlusions. In comparison, a set of co-occurrent local patterns can be less ambiguous. Along this direction, pairwise spatial constraints among local interest points have been used to enhance image registration [12]; various types of spatial contextual information have been used for object detection [10] and action classification [24]; and a grouplet representation has been developed to capture discriminative visual features and their spatial configurations for detecting the human-object-interaction scenes in images [39].

Motivated by the importance of incorporating audio information to help video concept classification, as well as the success of using structured visual features for image classification, in this paper, we propose an Audio-Visual Grouplet (AVG) representation. Each AVG contains a set of audio and visual codewords that have strong temporal correlations in videos. An audio-visual dictionary can be constructed to classify concepts using AVGs as building blocks. The AVGs capture not only the individual audio and visual features carried by the discrete audio and visual codewords, but also the temporal relations between audio and visual channels. By using the entire AVGs as building elements to represent videos, various concepts can be more robustly classified than using discrete audio and visual codewords. For example, as illustrated in Figure 2, The AVG that captures the visual bride and audio speech gives a strong audio-visual cue to classify the “wedding ceremony” concept, and the AVG that captures the visual bride and audio dancing music is quite discriminative to classify the “wedding dance” concept.

## 2. OVERVIEW OF OUR APPROACH

Figure 3 summarizes the framework of our system. We discover four types of AVGs by exploring four types of tem-

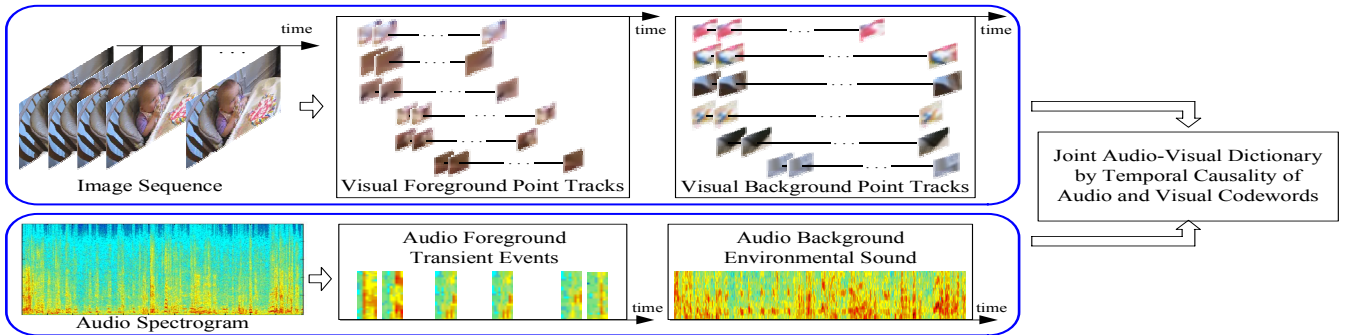
<sup>1</sup>Detailed definition of “non-music performance” can be found in [18].



**Figure 2: An example of AVG-based audio-visual dictionary.** Each AVG is composed of a set of audio and visual codewords that have strong temporal correlations in videos. The AVG that captures the visual bride and audio speech (AVG:  $w_1^{visual}, w_2^{visual}, w_1^{audio}$ ) gives a unique audio-visual cue to classify “wedding ceremony,” and the AVG that captures the visual bride and audio dancing music (AVG:  $w_1^{visual}, w_2^{visual}, w_2^{audio}$ ) is discriminative to classify “wedding dance.” In comparison, discrete visual or audio codewords can be ambiguous for classification.

poral audio-visual correlations: correlations between visual foreground and audio foreground; correlations between visual background and audio background; correlations between visual foreground and audio background; and correlations between visual background and audio foreground. All of these types of AVGs are useful for video concept classification. For example, as illustrated in Figure 3, to effectively classify the “birthday” concept, all of the following factors are important: the visual foreground people (*e.g.*, baby and child), the visual background setting (*e.g.*, cake and table), the audio foreground sound (*e.g.*, cheering, birthday song, and hand clapping), and the audio background sound (*e.g.*, music). By studying the temporal audio-visual correlations among these factors, we can identify unique audio-visual patterns that are discriminative for “birthday” classification.

To enable the exploration of the foreground and background audio-visual correlations, coarse-level separation of the foreground and background is needed in both visual and audio channels. It is worth mentioning that due to the diverse video content and the challenging conditions (*e.g.*, un-



**Figure 3:** The overall framework of the proposed joint audio-visual analysis system. The example shows a “birthday” video, where four types of audio-visual patterns are useful for classifying the “birthday” concept: (1) the visual foreground baby with the audio foreground events such as singing the birthday song or people cheering, since a major portion of birthday videos have babies or children involved; (2) the visual foreground baby with the audio background music; (3) the visual background setting such as the cake, with the audio foreground singing/cheering; and (4) the visual background cake with the audio background music.

even lighting, clutter, occlusions, complicated objects and camera motions, and the unstructured audio sounds with overlapping acoustic sources), precise visual or audio foreground/background separation is infeasible in generic videos. In addition, exact audio-visual synchronization can be unreliable most of the time. Multiple moving objects usually make sounds together, and often the object making sounds does not synchronically appear in video. To accommodate these issues, different from most previous audio-visual analysis methods [3, 7, 15, 31] that rely on precisely separated visual foreground objects and/or audio foreground sounds, our proposed approach has the following characteristics.

- We explore statistical temporal audio-visual correlations over a set of videos instead of exact audio-visual synchronization in individual videos. By representing the temporal sequences of visual and audio codewords as multivariate point processes, the statistical pairwise nonparametric Granger causality [14] between audio and visual codewords is analyzed. Based on the audio-visual causal matrix, salient AVGs are identified, which encapsulate strongly correlated visual and audio codewords as building blocks to classify videos.
- We do not pursue precise visual foreground/background separation. We aim to build foreground-oriented and background-oriented visual vocabularies. Specifically, consistent local points are tracked throughout each video. Based on both local motion vectors and spatiotemporal analysis of whole images, the point tracks are separated into foreground tracks and background tracks. Due to the challenging conditions of generic videos, such a separation is not precise. The target is to maintain a majority of foreground (background) tracks so that the constructed visual foreground (background) vocabulary can capture mainly visual foreground (background) information.
- Similar to the visual aspect, we aim to build foreground-oriented and background-oriented audio vocabularies, instead of pursuing precisely separated audio foreground or background acoustic sources. In generic videos, the foreground sound events are usually distributed unevenly and sparsely. Therefore, the local representation that focuses on short-term transient sound events [6] can be used to capture the foreground audio information. Also, the Mel-Frequency Cepstral Coefficients

(*MFCCs*) extracted from uniformly spaced audio windows roughly capture the overall information of the environmental sound. Based on the local representation and *MFCCs*, audio foreground and background vocabularies can be constructed, respectively.

We extensively evaluate our approach over the large-scale CCV set [18], containing 9317 consumer videos from YouTube. The consumer videos are captured by ordinary users under uncontrolled challenging conditions, without post-editing. The original audio soundtracks are preserved, which allows us to study legitimate audio-visual interactions. Experiments show that compared with the state-of-the-art multi-modal fusion methods using BoW representations, our AVG-based dictionaries can capture useful audio-visual cues and significantly improve the classification performance.

### 3. BRIEF REVIEW OF RELATED WORK

#### 3.1 Audio-Visual Concept Classification

Audio-visual analysis has been largely studied for speech recognition [15], speaker identification [31], and object localization [4]. For example, with multiple cameras and audio sensors, by using audio spatialization and multi-camera tracking, moving sound sources (*e.g.*, people) can be located. In videos captured by a single sensor, objects are usually located by studying the audio-visual synchronization along the temporal dimension. A common approach, for instance, is to project each of the audio and visual modalities into a 1D subspace and then correlate the 1D representations [3, 7]. These methods have shown interesting results in analyzing videos in a controlled or simple environment, where good sound source separation and visual foreground/background separation can be obtained. However, they cannot be easily applied to generic videos due to the difficulties in both acoustic source separation and visual object detection.

Most existing approaches for general video concept classification exploit the multi-modal fusion strategy instead of using direct correlation or synchronization across audio and visual modalities. For example, early fusion is used [5, 37] to concatenate features from different modalities into long vectors. This approach usually suffers from the “curse of dimensionality,” as the concatenated multi-modal feature can be very long. Also, it remains an open issue how to construct suitable joint feature vectors comprising features from

different modalities with different time scales and different distance metrics. In late fusion, individual classifiers are built for each modality separately, and their judgments are combined to make the final decision. Several combination strategies have been used, such as majority voting, linear combination, super-kernel nonlinear fusion [37], or SVM-based meta-classification combination [22]. However, classifier combination remains a basic machine learning problem. The best strategies depend on particular tasks.

Recently, an Audio-Visual Atom (AVA) representation has been developed in [16]. Visual regions are tracked within short-term video slices to generate visual atoms, and audio energy onsets are located to generate audio atoms. Regional visual features extracted from visual atoms and spectrogram features extracted from audio atoms are concatenated to form the AVA representation. An audio-visual codebook is constructed based on the AVAs for concept classification. The audio-visual synchrony is found through learning the audio-visual codebook. However, the temporal audio-visual interaction remains unstudied. As illustrated in Figure 1, the temporal audio-visual dependencies can reveal unique audio-visual patterns to assist concept classification.

### 3.2 Visual Foreground/Background Separation

One of the most commonly used techniques for separating foreground moving objects and the static background is background subtraction, where foreground objects are detected as the difference between the current frame and a reference image of the static background [11]. Various threshold adaptation methods [1] and adaptive background models [33] have been developed. However, these approaches require a relatively static camera, small illumination change, simple and stable background scene, and relatively slow object motion. When applied to generic videos for general concept classification, the performance is not satisfactory.

Motion-based segmentation methods have also been used to separate moving foreground and static background in videos [20]. The dense optical flow is usually computed to capture pixel-level motions. Due to the sensitivity to large camera/object motion and the computation intensity, such methods cannot be easily applied to generic videos either.

### 3.3 Audio Source Separation

Real-world audio signals are combinations of a number of independent sound sources, such as various instrumental sounds, human voices, natural sounds, *etc.* Ideally, one would like to recover each source signal. However, this task is very challenging in generic videos, because only a single audio channel is available, and realistic soundtracks have unrestricted content from an unknown number of unstructured, overlapping acoustic sources.

Early Blind Audio Source Separation (BASS) methods separate audio sources that are recorded with multiple microphones [28]. Later on, several approaches have been developed to separate single-channel audio, such as the factorial HMM methods [30] and the spectral decomposition methods [35]. Recently, the visual information has been incorporated to assist BASS [36], where the audio-video synchrony is used as side information. However, soundtracks studied by these methods are mostly mixtures of human voices or instrumental sounds with very limited background noise. When applied to generic videos, existing BASS methods cannot perform satisfactorily.

## 4. VISUAL PROCESS

We conduct SIFT point tracking within each video, based on which foreground-oriented and background-oriented temporal visual patterns are generated. The following details the processing stages.

### 4.1 Excluding Bad Video Segments

Video shot boundary detection and bad video segment elimination are general preprocessing steps for video analysis. Each raw video is segmented into several parts according to the detected shot boundaries with a single shot in each part. Next, segments with very large camera motion are excluded from analysis. It is worth mentioning that in our case, these steps can actually be skipped, because we process consumer videos that have a single long shot per video, and the SIFT point tracking can automatically exclude bad segments by generating few tracks over such segments. However, we still recommend these preprocessing steps to accommodate a large variety of generic videos.

### 4.2 SIFT-based Point Tracking

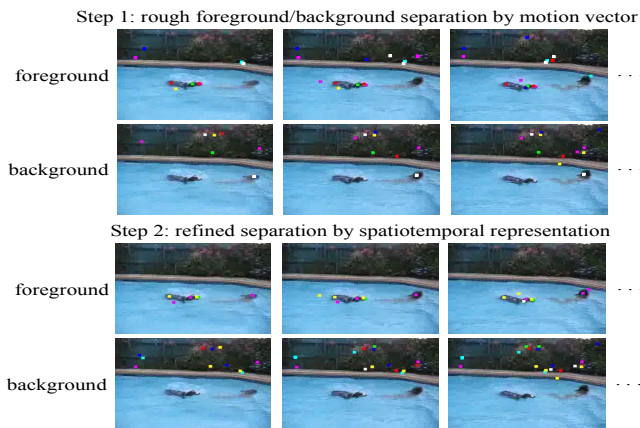
We use Lowe’s 128-dim SIFT descriptor with the DoG interest point detector [23]. SIFT features are first extracted from a set of uniformly sampled image frames with a sampling rate of 6 fps (frames per second)<sup>2</sup>. Then for adjacent image frames, pairs of matching SIFT features are found based on the Euclidean distance of their feature vectors, by also using Lowe’s method to discard ambiguous matches [23]. After that, along the temporal dimension, the matching pairs are connected into a set of SIFT point tracks, where different point tracks can start from different image frames and last variable lengths. This 6 fps sampling rate is empirically determined by considering both the computation cost and the ability of SIFT matching. In general, increasing the sampling rate will decrease the chance of missing point tracks, with the price of increased computation.

Each SIFT point track is represented by a 136-dim feature vector. This feature vector is composed by a 128-dim SIFT vector concatenated with an 8-dim motion vector. The SIFT vector is the averaged SIFT features of all SIFT points in the track. The motion vector is the averaged Histogram of Oriented Motion (HOM) along the track. That is, for each adjacent matching pair in the track, we compute the speed and direction of the local motion vector. By quantizing the 2D motion space into 8 bins (corresponding to 8 directions), an 8-dim HOM feature is computed where the value over each bin is the averaged speed of the motion vectors from the track moving along this direction.

### 4.3 Foreground/Background Separation

Once the set of SIFT point tracks are obtained, we separate them as foreground or background with the following two steps, as illustrated in Figure 4. First, for two adjacent frames  $I_i$  and  $I_{i+1}$ , we roughly separate their matching SIFT pairs into candidate foreground and background pairs based on the motion vectors. Specifically, we group these matching pairs by hierarchical clustering, where the grouping criterion is that pairs within a cluster have roughly the same moving direction and speed. Those SIFT pairs in the biggest cluster are treated as candidate background pairs, and all other

<sup>2</sup>In our experiment, the typical frame rate of videos is 30 fps. Typically we sample 1 frame from every 5 frames.



**Figure 4: Example of foreground/background separation of SIFT point tracks. A rough separation is obtained by analyzing local motion vectors. The result is further refined by spatiotemporal analysis over entire images.**

pairs are treated as candidate foreground pairs. The rationale is that foreground moving objects usually occupy less than half of the entire screen, and points on the foreground objects do not have a very consistent moving pattern. In comparison, points on the static background generally have consistent motion and this motion is caused by camera motion. This first step can distinguish background tracks fairly well for videos with moderate planar camera motions that occur most commonly in generic videos.

In the second step, we further refine the candidate foreground and background SIFT tracks by using the spatiotemporal representation of videos. A spatiotemporal X-ray image representation has been proposed by Akutsu and Tonomura for camera work identification [2], where the average of each line and each column in successive images are computed. The distribution of the angles of edges in the X-ray images can be matched to camera work models, from which camera motion classification and temporal video segmentation can be obtained directly [19]. When used alone, such methods cannot generate satisfactory segmentation results in many generic videos where large motions from multiple objects cannot be easily discriminated from the noisy background motion. The performance drops even more for small resolutions, *e.g.*,  $320 \times 240$  for most videos in our experiments. Therefore, instead of pursuing precise spatiotemporal object segmentation, we use such a spatiotemporal analysis to refine the candidate foreground and background SIFT tracks. The spatiotemporal image representation is able to capture camera zoom and tilt, which can be used to rectify those candidate tracks that are mistakenly labeled as foreground due to camera zoom and tilt. Figure 4 shows an example of visual foreground/background separation by using the above two steps.

#### 4.4 Vocabularies and Feature Representations

Based on the foreground and background SIFT point tracks, we build a visual foreground vocabulary and a visual background vocabulary, respectively. The BoW features can be computed using the vocabularies, which can be used directly for concept classification. Also, the temporal patterns of codeword occurrences can be computed to study the correlations between audio and visual signals in Section 6.

From Section 4.2, each SIFT point track is represented by a 136-dim feature vector. All foreground tracks from

the training videos are collected together, based on which the hierarchical K-means technique [27] is used to construct a  $D$ -word foreground visual vocabulary  $\mathcal{V}^{f-v}$ . Similarly, a  $D$ -word background visual vocabulary  $\mathcal{V}^{b-v}$  is constructed with all of the training background tracks. In our experiments, we use relatively large vocabularies,  $D=4000$ , based on findings from previous literatures [38] that when the vocabulary size exceeds 2000, the classification performance tends to saturate. That is, we can alleviate the influence of the vocabulary size on the final classification performance.

For each video  $V_j$ , all of its foreground SIFT point tracks are matched to the foreground codewords. A soft weighting scheme is used to alleviate the quantization effects [17], and a  $D$ -dim foreground BoW feature  $\mathbf{F}_j^{f-v}$  is generated. Similarly, all of the background SIFT point tracks are matched to the background codewords to generate a  $D$ -dim background BoW feature  $\mathbf{F}_j^{b-v}$ . In general, both  $\mathbf{F}_j^{f-v}$  and  $\mathbf{F}_j^{b-v}$  have their impacts in classifying concepts, *e.g.*, both the foreground people with caps and gowns and the background stadium setting are useful to classify “graduation” videos.

To study the temporal audio-visual interactions, the following histogram feature is computed over time for each of the foreground and background visual vocabularies. Given a video  $V_j$ , we have a set of foreground SIFT point tracks. Each track is labeled to one codeword in vocabulary  $\mathcal{V}^{f-v}$  that is closest to the track in the visual feature space. Next, for each frame  $I_{ji}$  in the video, we count the occurring frequency of each foreground codeword labeled to the foreground SIFT point tracks that have a SIFT point falling in this frame, and a  $D$ -dim histogram  $H_{ji}^{f-v}$  can be generated. Similarly, we can generate a  $D$ -dim histogram  $H_{ji}^{b-v}$  for each image frame  $I_{ji}$  based on vocabulary  $\mathcal{V}^{b-v}$ . After this computation, for the foreground  $\mathcal{V}^{f-v}$  (or background  $\mathcal{V}^{b-v}$ ), we have a temporal sequence  $\{H_{j1}^{f-v}, H_{j2}^{f-v}, \dots\}$  (or  $\{H_{j1}^{b-v}, H_{j2}^{b-v}, \dots\}$ ) over each video  $V_j$ .

## 5. AUDIO PROCESS

Instead of pursuing precisely separated audio sound sources, we extract background-oriented and foreground-oriented audio features. The temporal interactions of these features with their visual counterparts can be studied to generate useful audio-visual patterns for concept classification.

### 5.1 Audio Background

Various descriptors have been developed to represent audio signals in both temporal and spectral domains. Among these features, the MFCCs feature is one of the most popular choices for many different audio recognition systems [5, 31]. MFCCs represent the shape of the overall spectrum with a few coefficients, and have been shown to work well for both structured sounds (*e.g.*, speech) and unstructured environmental sounds. In soundtracks of generic videos, the foreground sound events (*e.g.*, an occasional dog barking or hand clapping) are distributed unevenly and sparsely. In such a case, the MFCCs extracted from uniformly spaced audio windows capture the overall characteristics of the background environmental sound, since the statistical impact of the sparse foreground sound events is quite small. Therefore, we use the MFCCs as the background audio feature.

For each given video  $V_j$ , we extract the 13-dim MFCCs from the corresponding soundtrack using 25 ms windows with a hop size of 10 ms. Next, we put all of the MFCCs

from all training videos together, on top of which the hierarchical K-means technique [27] is used to construct a  $D$ -word background audio vocabulary  $\mathcal{V}^{b-a}$ . Similar to visual-based processing, we compute two different histogram-like features based on  $\mathcal{V}^{b-a}$ . First, we generate a BoW feature  $\mathbf{F}_j^{b-a}$  for each video  $V_j$  by matching the MFCCs in the video to codewords in the vocabulary and conducting soft weighting. This BoW feature can be used directly for classifying concepts. Second, to study the audio-visual correlation, a temporal audio histogram sequence  $\{H_{j_1}^{b-a}, H_{j_2}^{b-a}, \dots\}$  is generated for each video  $V_j$  as follows. Each MFCC vector is labeled to one codeword in the audio background vocabulary  $\mathcal{V}^{b-a}$  that is closest to the MFCC vector. Next, for each sampled image frame  $I_{ji}$  in the video, we take a 200 ms window centered on this frame. Then we count the occurring frequency of the codewords labeled to the MFCCs that fall into this window, and a  $D$ -dim histogram  $H_{ji}^{b-a}$  can be generated. This  $H_{ji}^{b-a}$  can be considered as temporally synchronized with the visual-based histograms  $H_{ji}^{f-v}$  or  $H_{ji}^{b-v}$ .

## 5.2 Audio Foreground

As mentioned above, the soundtrack of a generic video usually has unevenly and sparsely distributed foreground sound events. To capture such foreground information, local representations that focus on short-term local sound events should be used. In [6], Cotton *et al.* have developed a local event-based representation, where a set of salient points in the soundtrack are located based on time-frequency energy analysis and multi-scale spectrum analysis. These salient points contain distinct event onsets, *i.e.*, transient events. By modeling the local temporal structure around each transient event, an audio feature reflecting the foreground of the soundtrack can be computed. In this work, we follow the recipe of [6] to generate the foreground audio feature.

Specifically, the Automatic Gain Control (AGC) is first applied to equalize the audio energy in both time and frequency domains. Next, the spectrogram of the AGC-equalized signal is taken for a number of different time-frequency trade-offs, corresponding to window length between 2 to 80 ms. Multiple scales enable the localization of events of different durations. High-magnitude bins in any spectrogram indicate a candidate transient event at the corresponding time. A limit is empirically set on the minimum distance between successive events to produce 4 events per second on average.

A 250 ms window of the audio signal is extracted centered on each transient event time, which captures the temporal structure of the transient event. Within each 250 ms window, a 40-dim spectrogram-based feature is computed for short-term signals over 25 ms windows with 10 ms hops, which results in 23 successive features for each event. These features are concatenated together to form a 920-dim representation for each transient event. After that, PCA is performed over all transient events from all training videos, and the top 20 bases are used to project the original 920-dim event representation to 20 dimensions.

By putting all the projected transient features from all training videos together, the hierarchical K-means technique [27] is used again to construct a  $D$ -word foreground audio vocabulary  $\mathcal{V}^{f-a}$ . We also compute two different histogram-like features based on  $\mathcal{V}^{f-a}$ . First, we generate a BoW feature  $\mathbf{F}_j^{f-a}$  for each video  $V_j$  by matching the transient features in the video to codewords in the vocabulary and conducting soft weighting. Second, a temporal audio histogram

sequence  $\{H_{j_1}^{f-a}, H_{j_2}^{f-a}, \dots\}$  is generated for each video  $V_j$  as follows. Each transient event is labeled to one codeword in the audio foreground vocabulary  $\mathcal{V}^{f-a}$  that is closest to the transient event feature. Next, for each sampled image frame  $I_{ji}$  in the video, we take a 200 ms window centered on this frame. Then we count the occurring frequency of the codewords labeled to the transient events whose centers fall into this window, and a  $D$ -dim histogram  $H_{ji}^{f-a}$  can be generated. Similar to  $H_{ji}^{b-a}$ ,  $H_{ji}^{f-a}$  can be considered as synchronized with  $H_{ji}^{f-v}$  or  $H_{ji}^{b-v}$ .

## 6. AVGS FROM TEMPORAL CAUSALITY

Recently, Prabhakar *et al.* have shown in [29] that the sequence of visual codewords produced by a space-time vocabulary representation of a video sequence can be interpreted as a multivariate point process. The pairwise temporal causal relations between visual codewords are computed within a video sequence, and visual codewords are grouped into causal sets. Evaluations over social game videos show promising results that the manually selected causal sets can capture the dyadic interactions. However, the work in [29] relies on nicely separated foreground objects, and causal sets are manually selected for each individual video. The method cannot be used for general concept classification.

We propose to investigate the temporal causal relations between audio and visual codewords. Our target is to capture the co-occurrences of audio and visual patterns over time. The rough foreground/background separation of both temporal SIFT point tracks and audio sounds enables a meaningful study of such temporal relations. For the purpose of classifying general concepts in generic videos, all of the following factors have their contributions: foreground visual objects, foreground audio transient events, background visual scenes, and background environmental sounds. Therefore, we explore their mixed-and-matched temporal relations to find salient AVGs that can assist the final classification.

### 6.1 Point-process Representation of Codewords

From the previous sections, for each video  $V_j$ , we have 4 temporal sequences:  $\{H_{j_1}^{f-v}, H_{j_2}^{f-v}, \dots\}$ ,  $\{H_{j_1}^{f-a}, H_{j_2}^{f-a}, \dots\}$ ,  $\{H_{j_1}^{b-v}, H_{j_2}^{b-v}, \dots\}$ , and  $\{H_{j_1}^{b-a}, H_{j_2}^{b-a}, \dots\}$ , with vocabularies  $\mathcal{V}^{f-v}$ ,  $\mathcal{V}^{f-a}$ ,  $\mathcal{V}^{b-v}$ , and  $\mathcal{V}^{b-a}$ , respectively. For each vocabulary, *e.g.*, the foreground visual vocabulary  $\mathcal{V}^{f-v}$ , each codeword  $w_k$  in the vocabulary can be treated as a point process,  $N_k^{f-v}(t)$ , which counts the number of occurrences of  $w_k$  in the interval  $(0, t]$ . The number of occurrences of  $w_k$  in a small interval  $dt$  is  $dN_k^{f-v}(t) = N_k^{f-v}(t+dt) - N_k^{f-v}(t)$ , and  $E\{dN_k^{f-v}(t)/dt\} = \lambda_k^{f-v}$  is the mean intensity. For theoretical and practical convenience, the zero-mean process is considered, and  $N_k^{f-v}(t)$  is assumed as wide-sense stationary, mixing, and orderly<sup>3</sup> [25]. Point processes generated by all  $D$  codewords of vocabulary  $\mathcal{V}^{f-v}$  form a  $D$ -dim multivariate point process  $\mathbf{N}^{f-v}(t) = (N_1^{f-v}(t), \dots, N_D^{f-v}(t))^T$ . Each video  $V_j$  gives one realization (trial) of  $\mathbf{N}^{f-v}(t)$  with counting vector  $(h_{j_1}^{f-v}(t), h_{j_2}^{f-v}(t), \dots, h_{j_D}^{f-v}(t))^T$ , where  $h_{j_k}^{f-v}(t)$  is the value over the  $k$ -th bin of the histogram  $H_{j_t}^{f-v}$ . Similarly,  $D$ -dim multivariate point processes  $\mathbf{N}^{f-a}(t)$ ,  $\mathbf{N}^{b-v}(t)$ ,

<sup>3</sup>The statistical properties of the process are time-invariant; the numbers of codewords in intervals separated widely in time are independent; given a sufficiently small interval, the number of codewords in the interval is at most one.

and  $\mathbf{N}^{b-a}(t)$  can be generated for vocabularies  $\mathcal{V}^{f-a}$ ,  $\mathcal{V}^{b-v}$ , and  $\mathcal{V}^{b-a}$ , respectively.

## 6.2 Temporal Causality among Codewords

Granger causality [14] is a statistical measure based on the concept of time series forecasting, where a time series  $Y_1$  is considered to causally influence a time series  $Y_2$  if predictions of future values of  $Y_2$  based on the joint history of  $Y_1$  and  $Y_2$  are more accurate than predictions based on  $Y_2$  alone. The estimation of Granger causality usually relies on autoregressive models, and for continuous-valued data like electroencephalogram, such model fitting is straightforward.

In [25], a nonparametric method that bypasses the autoregressive model fitting has been developed to estimate Granger causality for point processes. The theoretical basis lies in the spectral representation of point processes, the factorization of spectral matrices, and the formulation of Granger causality in the spectral domain. In the following, we describe the details of using the method of [25] to compute the temporal causality between audio and visual codewords. For simplicity, we temporarily omit indexes  $f-v$ ,  $b-v$ ,  $f-a$ , and  $b-a$ , w.l.o.g., since Granger causality can be computed for any two codewords from any vocabularies.

### 6.2.1 Spectral representation of point processes

The pairwise statistical relation between two point processes  $N_k(t)$  and  $N_l(t)$  can be captured by the cross-covariance density function  $R_{kl}(u)$  at lag  $u$ :

$$R_{kl}(u) = \frac{E\{dN_k(t+u)dN_l(t)\}}{du dt} - I[N_k(t) = N_l(t)] \lambda_k \delta(u),$$

where  $\delta(u)$  is the classical Kronecker delta function, and  $I[\cdot]$  is the indicator function. By taking the Fourier transform of  $R_{kl}(u)$ , we obtain the cross-spectrum  $S_{kl}(f)$ . Specifically, the multitaper method [34] can be used to compute the spectrum, where  $M$  data tapers  $\{q_m\}_{m=1}^M$  are applied successively to point process  $N_k(t)$  (with length  $T$ ):

$$S_{kl}(f) = \frac{1}{2\pi MT} \sum_{m=1}^M \tilde{N}_k(f, m) \tilde{N}_l(f, m)^*, \quad (1)$$

$$\tilde{N}_k(f, m) = \sum_{t_p=1}^T q_m(t_p) N_k(t_p) \exp(-2\pi i f t_p).$$

The symbol  $*$  is the complex conjugate transpose. Equation (1) gives an estimation of the cross-spectrum using one realization, and such estimations of multiple realizations are averaged to give the final estimation of the cross-spectrum.

### 6.2.2 Granger causality in spectral domain

For multivariate continuous-valued time series  $Y_1$  and  $Y_2$  with joint autoregressive representations:

$$Y_1(t) = \sum_{p=1}^{\infty} a_p Y_1(t-p) + \sum_{p=1}^{\infty} b_p Y_2(t-p) + \epsilon(t),$$

$$Y_2(t) = \sum_{p=1}^{\infty} c_p Y_2(t-p) + \sum_{p=1}^{\infty} d_p Y_1(t-p) + \eta(t),$$

their noise terms are uncorrelated over time and their contemporaneous covariance matrix is:

$$\Sigma = \begin{bmatrix} \Sigma_2 \Upsilon_2 \\ \Upsilon_2 \Gamma_2 \end{bmatrix}, \Sigma_2 = \text{var}(\epsilon(t)), \Gamma_2 = \text{var}(\eta(t)), \Upsilon_2 = \text{cov}(\epsilon(t), \eta(t)).$$

We can compute the spectral matrix as [9]:

$$\mathbf{S}(f) = \begin{bmatrix} S_{11}(f) & S_{12}(f) \\ S_{21}(f) & S_{22}(f) \end{bmatrix} = \mathbf{H}(f) \Sigma \mathbf{H}(f)^*, \quad (2)$$

where  $\mathbf{H}(f) = \begin{bmatrix} H_{11}(f) & H_{12}(f) \\ H_{21}(f) & H_{22}(f) \end{bmatrix}$  is the transfer function depending on coefficients of the autoregressive model. The

spectral matrix  $\mathbf{S}(f)$  of two point processes  $N_k(t)$  and  $N_l(t)$  can be estimated using Eqn. (1). By spectral matrix factorization we can decompose  $\mathbf{S}(f)$  into a unique corresponding transfer function  $\tilde{\mathbf{H}}(f)$  and noise processes  $\tilde{\Sigma}_2$  and  $\tilde{\Gamma}_2$ . Next, the Granger causality at frequency  $f$  can be estimated according to the algorithm developed in [9]:

$$G_{N_l \rightarrow N_k}(f) = \ln \left( \frac{S_{kk}(f)}{\tilde{H}_{kk}(f) \tilde{\Sigma}_2 \tilde{H}_{kk}(f)^*} \right), \quad (3)$$

$$G_{N_k \rightarrow N_l}(f) = \ln \left( \frac{S_{ll}(f)}{\tilde{H}_{ll}(f) \tilde{\Gamma}_2 \tilde{H}_{ll}(f)^*} \right). \quad (4)$$

The Granger causality scores over all frequencies are then summed together to obtain a single time-domain causal influence, *i.e.*,  $C_{N_k \rightarrow N_l} = \sum_f G_{N_k \rightarrow N_l}(f)$ , and  $C_{N_l \rightarrow N_k} = \sum_f G_{N_l \rightarrow N_k}(f)$ . In general,  $C_{N_k \rightarrow N_l} \neq C_{N_l \rightarrow N_k}$ , due to the directionality of the causal relations.

## 6.3 AVGs from the Causal Matrix

Our target of studying temporal causality between audio and visual codewords is to identify strongly correlated AVGs, where the direction of the relations is usually not important. For example, a dog can start barking at any time during the video, and we would like to find the AVG that contains correlated codewords describing the foreground dog barking sound and the visual dog point tracks. The direction of whether the barking sound is captured before or after the visual tracks is irrelevant. Therefore, for a pair of codewords represented by point processes  $N_k^{s_k}(t)$  and  $N_l^{s_l}(t)$  (where  $s_k$  or  $s_l$  is one of the following  $f-v$ ,  $f-a$ ,  $b-v$ , and  $b-a$ , indicating the vocabularies the codeword comes from), the nonparametric Granger causality scores from both directions  $C_{N_k^{s_k} \rightarrow N_l^{s_l}}$  and  $C_{N_l^{s_l} \rightarrow N_k^{s_k}}$  are summed together to generate the final similarity between these two codewords:

$$C(N_k^{s_k}, N_l^{s_l}) = C_{N_k^{s_k} \rightarrow N_l^{s_l}} + C_{N_l^{s_l} \rightarrow N_k^{s_k}}. \quad (5)$$

Then, for a pair of audio and visual vocabularies, *e.g.*,  $\mathcal{V}^{f-v}$  and  $\mathcal{V}^{f-a}$ , we have a  $2D \times 2D$  symmetric causal matrix:

$$\begin{bmatrix} \mathbf{C}^{f-v, f-v} & \mathbf{C}^{f-v, f-a} \\ \mathbf{C}^{f-a, f-v} & \mathbf{C}^{f-a, f-a} \end{bmatrix}, \quad (6)$$

where  $\mathbf{C}^{f-v, f-v}$ ,  $\mathbf{C}^{f-a, f-a}$ , and  $\mathbf{C}^{f-v, f-a}$  are  $D \times D$  matrices with entries  $C(N_k^{f-v}, N_l^{f-v})$ ,  $C(N_k^{f-a}, N_l^{f-a})$ , and  $C(N_k^{f-v}, N_l^{f-a})$ , respectively.

Spectral clustering can be applied directly based on this causal matrix to identify groups of codewords that have high correlations. Here we use the algorithm developed in [26] where the number of clusters can be determined automatically by analyzing the eigenvalues of the causal matrix. Each cluster is called an AVG, and codewords in an AVG can come from both audio and visual vocabularies. The AVGs capture temporally correlated audio and visual codewords that statistically interact over time. Each AVG can be treated as an audio-visual pattern, and all AVGs form an audio-visual dictionary. For a given video  $V_j$ , the original SIFT point tracks and audio features can be mapped to AVGs and generate an audio-visual dictionary-based feature. The simplest way to compute this dictionary-based feature is to aggregate the original BoW features over individual visual and audio vocabularies. For instance, for an AVG containing codewords  $w_1^{f-a}, \dots, w_n^{f-a}$  from  $\mathcal{V}^{f-a}$ , and  $w_1^{f-v}, \dots, w_m^{f-v}$  from  $\mathcal{V}^{f-v}$ , the value over the corresponding bin in the dictionary-based feature of video  $V_j$  is:  $\sum_{p=1}^n F_{jp}^{f-a} + \sum_{p=1}^m F_{jp}^{f-v}$ .  $F_{jp}^{f-a}$  ( $F_{jp}^{f-v}$ ) is the value over the  $p$ -th bin in the BoW feature

$\mathbf{F}_j^{f-a}$  ( $\mathbf{F}_j^{f-v}$ ) over  $\mathcal{V}^{f-a}$  ( $\mathcal{V}^{f-v}$ ). Classifiers such as SVMs can be trained using this feature for concept classification.

A total of four audio-visual dictionaries are generated in this work, by studying the temporal causal relations between different types of audio and visual codewords. They are: dictionary  $\mathcal{D}^{f-v, f-a}$  by correlating  $\mathcal{V}^{f-v}$  and  $\mathcal{V}^{f-a}$ ,  $\mathcal{D}^{b-v, b-a}$  by correlating  $\mathcal{V}^{b-v}$  and  $\mathcal{V}^{b-a}$ ,  $\mathcal{D}^{f-v, b-a}$  by correlating  $\mathcal{V}^{f-v}$  and  $\mathcal{V}^{b-a}$ , and  $\mathcal{D}^{b-v, f-a}$  by correlating  $\mathcal{V}^{b-v}$  and  $\mathcal{V}^{f-a}$ . As illustrated in Figure 3, all of these correlations reveal useful audio-visual patterns for classifying concepts.

## 7. EXPERIMENTS

We evaluate our algorithm over the large-scale CCV set [18], containing 9317 consumer videos from YouTube. These videos are captured by ordinary users under unrestricted challenging conditions, without post-editing. The original audio soundtracks are preserved, in contrast to other large-scale news or movie video sets [21, 32]. This allows us to study legitimate audio-visual interactions. Each video is manually labeled to 20 semantic concepts by using Amazon Mechanical Turk. More details about the data set and category definitions can be found in [18]. Our experiments take similar settings as [18], *i.e.*, we use the same training (4659 videos) and test (4658 videos) sets, and one-versus-all  $\chi^2$ -kernel SVM classifiers. The performance is measured by Average Precision (*AP*, the area under uninterpolated PR curve) and Mean AP (*MAP*, averaged AP across concepts).

### 7.1 Performance of Concept Classification

To demonstrate the effectiveness of our method, we first evaluate the performance of the state-of-the-art BoW representations using different types of individual audio and visual features exploited in this paper, as well as the performance of their various early-fusion combinations. The AP and MAP results are shown in Figure 5. These BoW representations are generated using the same method as [18]. The results show that the individual visual SIFT, audio MFCCs, and audio transient event feature perform comparably overall, each having different advantages over different concepts. The combinations of audio and visual BoW representations through multi-modal fusion can consistently and significantly improve classification. For example, by combining the three individual features (“SIFT+MFCCs+Trans”), compared with individual features, all concepts get AP improvements, and the MAP is improved by over 33% on a relative basis. Readers may notice that our “SIFT” performs differently than that in [18]. This is because we have only a single type of SIFT feature (*i.e.*, SIFT over DoG keypoints) and generate the BoW representation using only the  $1 \times 1$  spatial layout, while several types of keypoints and spatial layouts are used in [18]. Actually, our “MFCCs” performs similarly to that in [18], due to the similar settings for feature extraction and vocabulary construction.

Next, we compare the classification performance of using individual foreground and background audio and visual vocabularies (*i.e.*,  $\mathcal{V}^{f-v}$ ,  $\mathcal{V}^{f-a}$ ,  $\mathcal{V}^{b-v}$ , and  $\mathcal{V}^{b-a}$ ), as well as using various types of audio-visual dictionaries (*i.e.*,  $\mathcal{D}^{f-v, f-a}$ ,  $\mathcal{D}^{f-v, b-a}$ ,  $\mathcal{D}^{b-v, f-a}$ , and  $\mathcal{D}^{b-v, b-a}$ ). Each audio-visual dictionary contains about 300 AVGs on average. The results are shown in Figure 6. Also, the four types of audio-visual-dictionary-based features are concatenated together to train concept classifiers so that the advantages of all dictionaries in classifying different concepts can be exploited. This is the

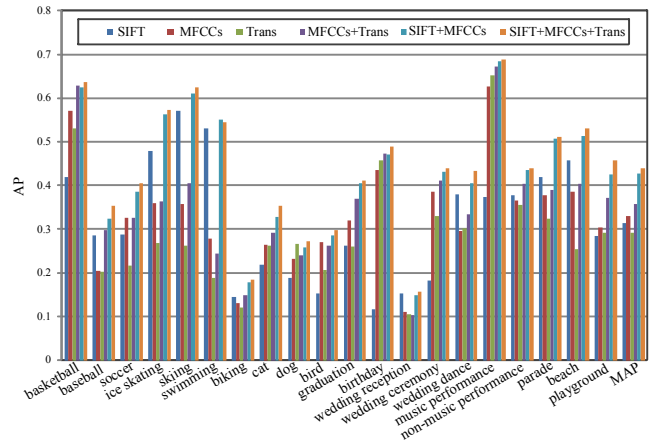


Figure 5: Comparison of various BoW representations as well as their early-fusion combinations.

“All A-V Dictionaries” approach shown in Figure 6. From the figure, we can see that for individual vocabularies, visual foreground performs better than visual background in general, while audio background performs better than audio foreground. Such results are within our expectation, because of the importance of the visual foreground in classifying objects and activities, as well as the effectiveness of audio background environmental sounds in classifying general concepts as shown by previous work [5, 18]. Compared with the visual foreground, visual background wins over “wedding ceremony” and “non-music performance,” because of the importance of the background settings for these concepts, *e.g.*, the arch, flower boutique, and seated crowd for “wedding ceremony,” and the stadium or stage setting for “non-music performance.” In the audio aspect, audio foreground outperforms audio background over three concepts, “dog,” “birthday,” and “music performance,” because of the usefulness of capturing consistent foreground sounds in these concepts.

Through exploring the temporal audio-visual interaction, the audio-visual dictionaries generally outperform the corresponding individual audio or visual vocabularies. For example, the MAP of  $\mathcal{D}^{f-v, f-a}$  outperforms those of  $\mathcal{V}^{f-v}$  and  $\mathcal{V}^{f-a}$  by more than 10%, the MAP of  $\mathcal{D}^{b-v, b-a}$  outperforms those of  $\mathcal{V}^{b-v}$  and  $\mathcal{V}^{b-a}$  by more than 20%, *etc.* By combining all types of dictionaries together, compared with individual audio or visual vocabularies, all concepts get consistent, significant AP improvements, and the overall MAP is improved by more than 50%. In addition, compared with direct multi-modal fusion without temporal audio-visual interaction (*i.e.*, “SIFT+MFCCs+Trans”), our “All A-V Dictionaries” approach has consistent AP gains over all concepts, with a 12% MAP gain overall. For 12 concepts (*e.g.*, “baseball,” “ice skating,” *etc.*) the APs are improved by more than 10%. The results demonstrate the effectiveness of extracting useful AVGs to classify general video concepts.

### 7.2 Visualization of AVGs

In Figure 7, we show some examples of the AVGs that are identified as carrying unique audio-visual patterns to classify different concepts. Such visualization helps us to subjectively and intuitively evaluate the effectiveness of our approach. Each example shows some visual and audio codewords grouped together in an AVG. Each visual codeword is shown as some local points in a video that are very similar to the codeword. For each audio codeword, we show the



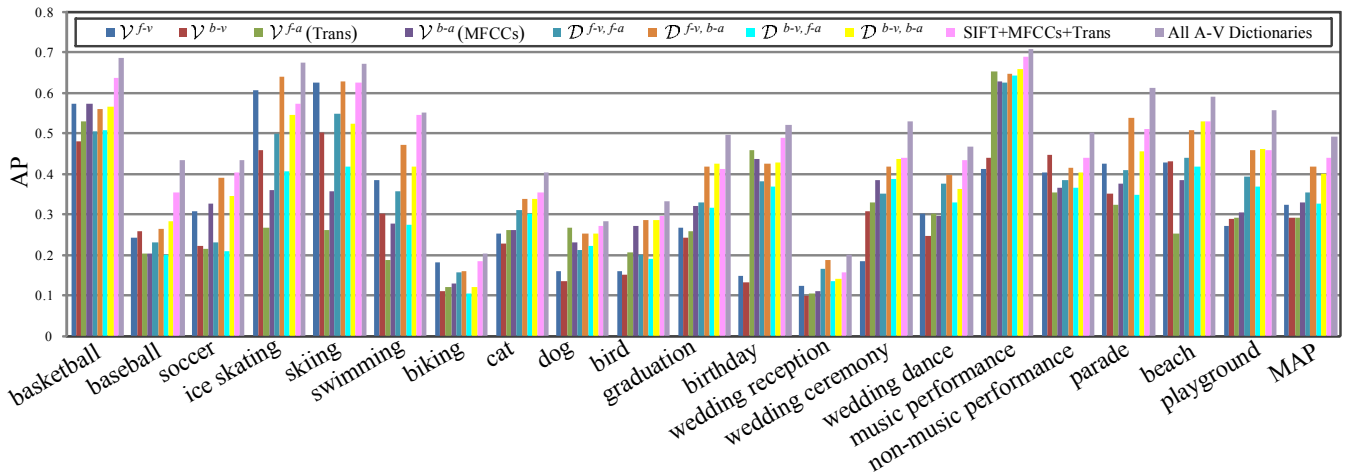


Figure 6: Comparison of individual foreground/background audio/visual vocabularies and audio-visual dictionaries.

spectrogram of the soundtrack over a 200 ms window, where the audio feature in this short window is very similar to the codeword. For the clarity of visualization, for each AVG, we show only a few codewords that co-occur in the same video. An actual AVG typically has about 20~30 codewords.

From the visualization we can see that by studying the temporal interactions of foreground and background audio and visual codewords, we can group inter-related audio and visual codewords into meaningful AVGs. Such AVGs provide discriminative audio-visual cues for classifying various concepts. The examples also confirm our assumption that all of the four types of audio-visual foreground/background interactions are important. For example, the visual foreground basketball player with audio foreground whistling and cheering sounds give a unique cue to classify “basketball” videos; the visual foreground bride and audio background music form a useful pattern to classify “wedding dance” videos; the visual background playground equipments and audio foreground children’s cheering and laughing sounds are useful to classify “playground” videos; and the visual background sand and water together with audio background wind blowing sound are discriminative to classify “beach” videos.

## 8. CONCLUSION

We proposed an AVG representation by studying the statistical temporal causality between audio and visual codewords. Each AVG encapsulates inter-related audio and visual codewords as a whole package, which carries unique audio-visual patterns to represent the video content. By conducting coarse-level foreground/background separation in both visual and audio channels, four types of AVGs were extracted from four types of temporal audio-visual correlations, correlations between visual foreground and audio foreground codewords, between visual foreground and audio background codewords, between visual background and audio foreground codewords, and between visual background and audio background codewords. Experiments over large-scale consumer videos demonstrate that all four types of AVGs provide discriminative audio-visual cues to classify various concepts, and significant performance improvements can be obtained compared with state-of-the-art multi-modal fusion methods using BoW representations. It is worth mentioning that our method has some limitations. For videos that we cannot get meaningful SIFT tracks or extract meaningful audio transient events, our method will not work well.

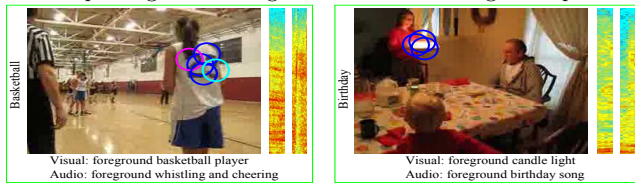
## 9. ACKNOWLEDGEMENT

We would like to thank the authors of [6] and [25] for sharing their code with us, and for Shih-Fu Chang for many useful discussions.

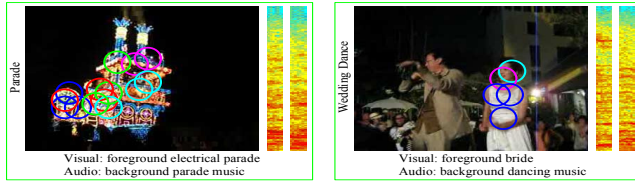
## 10. REFERENCES

- [1] T. Aach and A. Kaup. Bayesian algorithms for adaptive change detection in image sequences using Markov random fields. *Signal Processing: Image Communication*, vol. 7, pages 147-160, 1995.
- [2] A. Akutsu and Y. Tonomura. Video tomography: An efficient method for camerawork extraction and motion analysis. *ACM Multimedia*, 1994.
- [3] Z. Barzelay and Y. Schechner. Harmony in motion. In *Proc. IEEE CVPR*, pages 1-8, 2007.
- [4] M.J. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Trans. PAMI*, 25(7):828-836, 2003.
- [5] S. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. Loui, and J. Luo. Large-scale multimodal semantic concept detection for consumer video. In *Proc. ACM MIR*, pages 255-264, 2007.
- [6] C. Cotton, D. Ellis, and A. Loui. Soundtrack classification by transient events. In *IEEE ICASSP*, Prague, Czech Republic, 2011.
- [7] M. Cristani, B. Manuele, and M. Vittorio. Audio-visual event recognition in surveillance video sequences. *IEEE Trans. Multimedia*, 9(2):257-267, 2007.
- [8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE CVPR*, pages 886-893, 2005.
- [9] M. Ding, Y. Chen, and S.L. Bressler. Granger causality: Basic theory and applications to neuroscience. In *S. Scheluter, N. Winterhalder, J. Timmer, Handbook of Time Series Analysis*, Wiley, Weinheim, 2006.
- [10] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. In *IEEE CVPR*, Miami, Florida, 2009.
- [11] S.Y. Elhabian, K.M. El-Sayed. Moving object detection in spatial domain using background removal techniques - State-of-art. *Recent Patents on Computer Science*, 1(1):32-54, 2008.
- [12] O. Enqvist, K. Josephson, and F. Kahl. Optimal

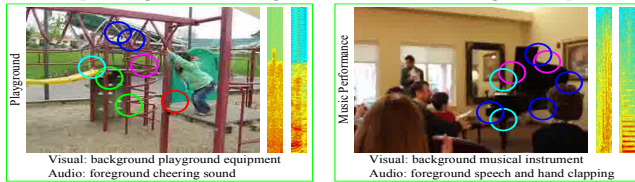
AVGs capturing visual foreground and audio foreground patterns.



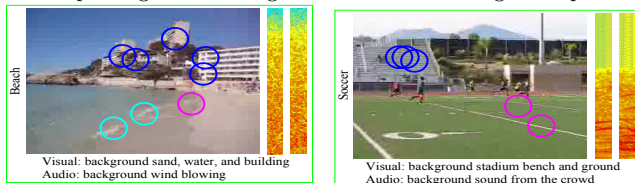
AVGs capturing visual foreground and audio background patterns.



AVGs capturing visual background and audio foreground patterns.



AVGs capturing visual background and audio background patterns.



**Figure 7: Examples of AVGs carrying useful audio-visual cues to classify concepts. For each AVG, we only show a few codewords that co-occur in an example video. For each video, visual codewords are marked by colored circles in a sampled image frame, each color indicating one visual codeword. Audio codewords are shown as the spectrograms of soundtracks over 200 ms windows. Only 2 audio codewords are given for each AVG. For instance, in the first example “basketball” video, we show 3 visual codewords and 2 audio codewords in an AVG.**

correspondences from pairwise constraints. In *IEEE ICCV*, Kyoto, Japan, 2009.

- [13] M. Everingham, L. Van Cool, C.K.I. Williams, J. Winn, and A. Zisserman. The Pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303-338, 2010.
- [14] C. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424-438, 1969.
- [15] K. Iwano, T. Yoshinaga, S. Tamura, and S. Furui. Audio-visual speech recognition using lip information extracted from side-face images. *EURASIP Journal on ASMP*, 2007(1):4-12, 2007.
- [16] W. Jiang, C. Cotton, S. Chang, D. Ellis, and A. Loui. Audio-visual atoms for generic video concept classification. *ACM TOMCCAP*, 6:1-19, 2010.
- [17] Y. Jiang, C.W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *ACM CIVR*, pages 494-501, 2007.
- [18] Y. Jiang, G. Ye, S. Chang, D. Ellis, and A. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *ACM ICMR*, Trento, Italy, 2011.

- [19] P. Joly and H.K. Kim. Efficient automatic analysis of camera work and microsegmentation of video using spatiotemporal images. *Signal Processing: Image Communication*, 8(4):295-307, 1996.
- [20] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *IEEE ICCV*, Brazil, 2007.
- [21] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE CVPR*, Alaska, 2008.
- [22] W.H. Lin and A. Hauptmann. News video classification using svm-based multimodal classifiers and combination strategies. In *Proc. ACM Multimedia*, pages 323-326, 2002.
- [23] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91-110, 2004.
- [24] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *IEEE CVPR*, Miami, Florida, 2009.
- [25] A. Nedungadi, G. Rangarajan, N. Jain, and M. Ding. Analyzing multiple spike trains with nonparametric granger causality. *Journal of Computational Neuroscience*, 27(1):55-64, 2009.
- [26] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2001.
- [27] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *Proc. IEEE CVPR*, pages 2161-2168, 2006.
- [28] D.T. Pham and J.F. Cardoso. Blind separation of instantaneous mixtures of non stationary sources. *IEEE Trans. Signal Processing*, 49(9):1837-1848, 2001.
- [29] K. Prabhakar, S. Oh, P. Wang, G. Abowd, and J.M. Rehg. Temporal causality for the analysis of visual events. In *IEEE CVPR*, San Francisco, 2010.
- [30] S.T. Roweis. One microphone source separation. In *NIPS*, pages 793-799, 2001.
- [31] M. Sargin, H. Aradhye, P.J. Moreno, and M. Zhao. Audiovisual celebrity recognition in unconstrained web videos. In *IEEE ICASSP*, Taipei, 2009.
- [32] A.F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVID. In *Proc. ACM MIR*, pages 321-330, 2006.
- [33] C. Stauffer and E. Grimson. Learning patterns of activity using realtime tracking. *IEEE Trans. PAMI*, 22(8):747-757, 2000.
- [34] A. Walden. A unified view of multitaper multivariate spectral estimation. *Biometrika*, 87(4):767-788, 2000.
- [35] B. Wang and M.D. Plumbley. Investigating single-channel audio source separation methods based on non-negative matrix factorization. In *Proc. ICARN*, pages 17-20, 2006.
- [36] W. Wang, D. Cosker, Y. Hicks, S. Saneit, and J. Chambers. Video assisted speech source separation. In *Proc. IEEE ICASSP*, pages 425-428, 2005.
- [37] Y. Wu, C.Y. Lin, E.Y. Chang, and J.R. Smith. Multimodal information fusion for video concept detection. In *Proc. IEEE ICIP*, pages 2391-2394, 2004.
- [38] J. Yang, Y. Jiang, A. Hauptmann, and C. Ngo. Evaluating bag-of-visual-words representations in scene classification. *ACM MIR*, pages 197-206, 2007.
- [39] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *IEEE CVPR*, San Francisco, 2010.