Int J Multimed Info Retr (2012) 1:223–238
DOI 10.1007/s13735-012-0020-6

REGULAR PAPER

Video concept detection by audio-visual grouplets

Wei Jiang · Alexander C. Loui

Received: 31 July 2012 / Accepted: 16 August 2012 / Published online: 7 September 2012

Abstract We investigate general concept classification in unconstrained videos by joint audio-visual analysis. An audio-visual grouplet (AVG) representation is proposed based on analyzing the statistical temporal audio-visual interactions. Each AVG contains a set of audio and visual codewords that are grouped together according to their strong temporal correlations in videos, and the AVG carries unique audio-visual cues to represent the video content. By using the entire AVGs as building elements, video concepts can be more robustly classified than using traditional vocabularies with 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. To effectively use the AVGs for improved concept classification, a distance metric learning algorithm is further developed. Based on the AVG structure, the algorithm uses an iterative quadratic programming formulation to learn the optimal distances between data points according to the large-margin nearest-neighbor setting. Various types of grouplet-based distances can be computed using individual AVGs, and through our distance metric learning algorithm these grouplet-based distances can be aggregated for final classification. We extensively evaluate our method over the large-scale Columbia consumer video set. Experiments demonstrate that the AVG-based audio-visual representation can achieve consistent and significant performance improvements compared wth other state-of-the-art approaches.

Keywords Video concept detection · Audio-visual grouplet

W. Jiang (✉) · A. C. Loui
Kodak Technology Center, Eastman Kodak Company,
Rochester, NY, USA
e-mail: wei.jiang@kodak.com

A. C. Loui
e-mail: alexander.loui@kodak.com

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 the TRECVID high-level feature extraction or multimedia event detection [34], the human action recognition in Hollywood movies [22], and the Columbia consumer video (CCV) concept detection [19]. 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 [24] or HOG [8]) are considered state-of-the-art. In a

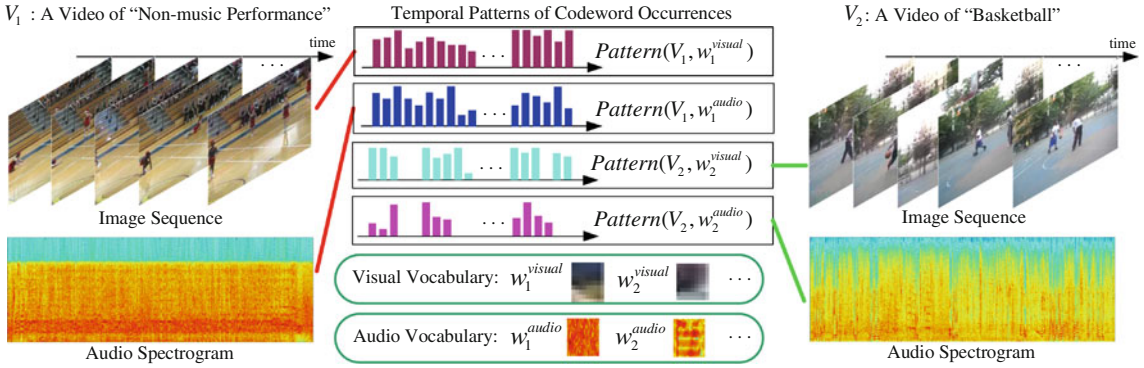


Fig. 1 Discovery of audio-visual patterns through temporal audio-visual interactions. w_i^{visual} and w_j^{audio} are discrete codewords in visual and audio vocabularies, respectively. By analyzing correlations between the temporal histograms of audio and visual codewords, we can discover salient audio-visual 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." The correlated visual stadium patches and audio background music are helpful to classify "non-music performance." In comparison, discrete audio and visual codewords are less discriminative than such audio-visual cues

BoW-based approach, local descriptors are vector-quantized against a vocabulary of prototypical descriptors to generate a histogram-like representation.

The importance of incorporating audio information to facilitate semantic concept classification has been discovered by several previous works [5,19,43]. They generally use a multi-modal fusion strategy, i.e., early fusion [5,19,43] to train classifiers with concatenated audio and visual features, or late fusion [5,43] to combine judgments from classifiers built over individual modalities. Different from such fusion approaches that avoid studying temporal audio-visual synchrony, the work in [17] 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 Fig. 1, by studying correlations between temporal patterns of visual and audio codewords, we can discover discriminative audio-visual cues, such as the encapsulation of visual basketball patches and audio basketball bouncing sounds for classifying "basketball," and the encapsulation of visual stadium patches and audio music sounds for classifying "non-music performance". To the best of our knowledge, such audio-visual cues have not been studied before in previous literature.

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 [13]; various types of spatial contextual information have been used for object detection [11,41] and action classification [25]; 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 [45].

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 Fig. 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.

In addition, we develop a distance metric learning algorithm to effectively use the extracted AVGs for classifying concepts. Based on the AVGs, an iterative Quadratic Programming (QP) problem is formulated to learn the optimal distance metric between data points based on the large-margin nearest neighbor (LMNN) setting [40]. Our distance metric learning framework is quite flexible, where various

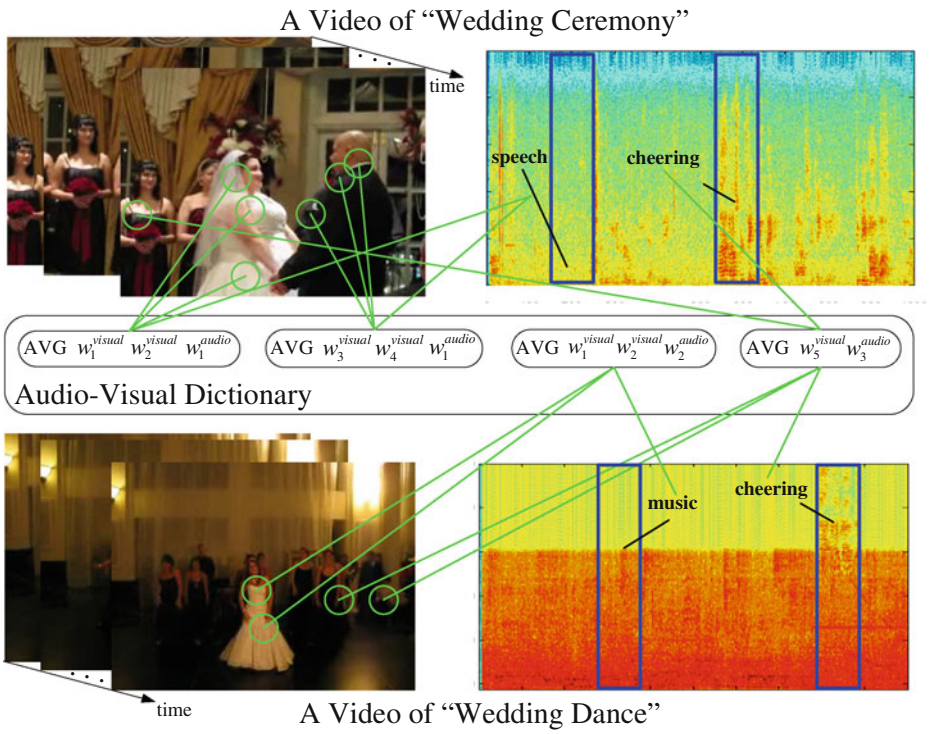


Fig. 2 An example of AVG-based audio-visual dictionary. Each AVG is composed of a set of audio and visual codewords that have strong temporal correlations in videos. The AVG that captures the visual bride and audio speech (AVG: w_1^{visual}, w_2^{visual}, w_1^{audio}) gives a unique audio-visual cue 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

types of grouplet-based distances can be computed using individual AVGs, and these grouplet-based distances can be fed into the same distance metric learning algorithm for concept classification. Specifically, we propose a grouplet-based distance based on the chi-square distance and word specificity [26], and through our distance metric learning such a grouplet-based distance can achieve consistent and significant classification performance gain.

2 Overview of our approach

Figure 3 summarizes the framework of our system. We discover four types of AVGs by exploring four types of temporal 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 Fig. 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., uneven lighting, clutter, occlusions, complicated objects and camera motions, and the unstructured audio sounds with overlapping acoustic sources), precise separation of visual or audio foreground and background 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 [3,7,16,32] 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 [15] 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.

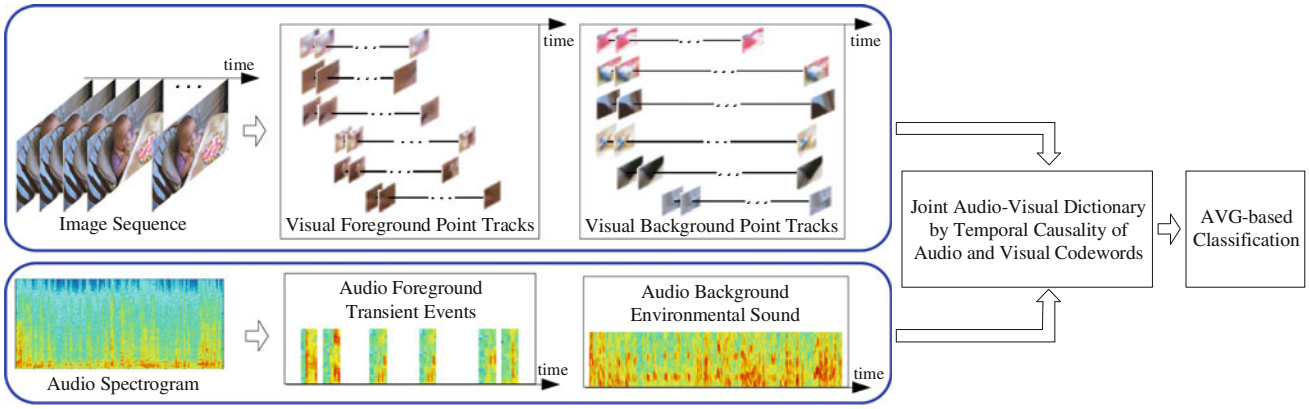


Fig. 3 The overall framework of the proposed joint audio-visual analysis system. The example shows a "birthday" video, where four types of audio-visual patterns are useful for classifying the "birthday" concept: (1) the visual foreground baby with the audio foreground events such as singing the happy 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

– 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 built, respectively.

After obtaining various types of AVGs, a distance metric learning algorithm is further developed to effectively use the AVGs for concept classification. Based on the AVGs, we learn the optimal distance metric between data points under the LMNN setting. LMNN is used because of its resemblance to SVMs, i.e., the role of large margin in LMNN is inspired by its role in SVMs, and LMNN should inherit various strengths of SVMs [33]. Therefore, the final learned distance metric can provide reasonably good performance for SVM concept classifiers.

We extensively evaluate our approach over the large-scale CCV set [19], 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 [16], speaker identification [32], 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 can not 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 [5,43] 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 the majority voting, linear combination, super-kernel nonlinear fusion [43], or SVM-based meta-classification combination [23]. However, effective classifier combination remains a basic machine learning problem. Recently, an audio-visual atom (AVA) representation has been developed in [17]. 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. The audio-visual synchrony is found through learning an audio-visual codebook based on the AVAs. However, the temporal audio-visual interaction remains unstudied. As illustrated in Fig. 1, the temporal audio-visual dependencies can reveal unique audio-visual patterns to assist concept classification. In addition, the work of [17] requires segmenting image frames into visual regions, which is too expensive to be practical.

3.2 Visual foreground/background separation

One most commonly used technique 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 [12]. Various threshold adaptation methods [1] and adaptive background models [35] have been developed. However, these approaches require a relatively static camera, small illumination change, simple and stable background scene, and relatively slow object motion. Their performances over generic videos are still not satisfactory.

Motion-based segmentation methods have also been used to separate moving foreground and static background in videos [21]. 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 human voices, instrumental sounds, 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 [29]. Later on, several approaches have been developed to separate single-channel audio, such as the factorial HMM methods [31] and the spectral decomposition methods [38]. Recently, the visual information has been incorporated to assist BASS [39], 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.

3.4 Distance metric learning

Distance metric learning is an important machine learning technique of adapting the underlying distance metric according to the available data for improved classification. The most popular distance metric learning algorithms are based on the Mahalanobis distance metric, such as the LMNN [40] method, the maximally collapsing metric learning approach [14], the information-theoretic metric learning method [9], and the semantic preserving BoW method [42]. However, it is non-trivial to incorporate the grouplet structure into the existing distance metric learning algorithms.

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 [24]. SIFT features are first extracted from a set of uniformly sampled image frames with a sampling rate of 6 fps (frames per second).[1] 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 [24]. 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 Fig. 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 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.

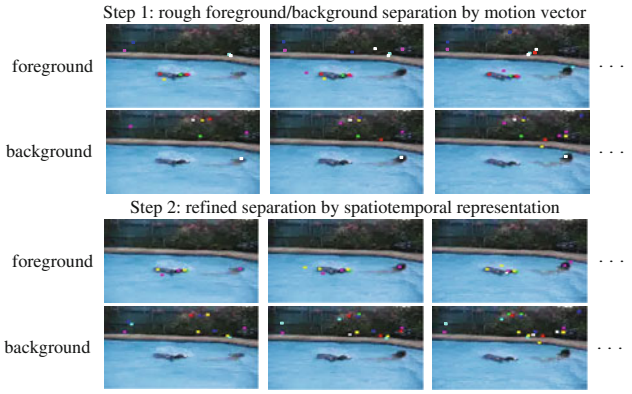


Fig. 4 Example of separating foreground/background SIFT tracks. A rough separation is obtained by analyzing local motion vectors. The result is further refined by spatiotemporal analysis over entire images

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 [20]. 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., 320×240 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 back-

[1] In our experiment, the typical frame rate of videos is 30 fps. Typically we sample 1 frame from every 5 frames.

ground vocabulary, respectively. The BoW features can be computed using the vocabularies, which can be used directly to classify concepts. Also, temporal patterns of codeword occurrences can be computed to study correlations between audio and visual signals in Sect. 6.

From Sect. 4.2, each SIFT 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 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 the previous literature [44] that when the vocabulary size exceeds 2000 the classification performance tends to saturate, we can alleviate the influence of the vocabulary size on the final classification performance. This size is also a tradeoff between accuracy and computational complexity.

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 [18], 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 $\{H_{j1}^{f-v}, H_{j2}^{f-v}, \ldots\}$ (or $\{H_{j1}^{b-v}, H_{j2}^{b-v}, \ldots\}$) 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 [5,32]. 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 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 $\{H_{j1}^{b-a}, H_{j2}^{b-a}, \ldots\}$ 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 tradeoffs, corresponding to window length between 2 and 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 four 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 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 $\{H_{j1}^{f-a}, H_{j2}^{f-a}, \ldots\}$ 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. [30] have shown 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 [30] 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. The rough separation of foreground and background for both temporal SIFT 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: $\{H_{j1}^{f-v}, H_{j2}^{f-v}, \ldots\}$, $\{H_{j1}^{f-a}, H_{j2}^{f-a}, \ldots\}$, $\{H_{j1}^{b-v}, H_{j2}^{b-v}, \ldots\}$, and $\{H_{j1}^{b-a}, H_{j2}^{b-a}, \ldots\}$, according to 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 [27]. Point processes generated by all D codewords of vocabulary $\mathcal{V}^{f-v}$ form a D-dim multivariate point process $\mathbf{N}^{f-v}(t) = (N_1^{f-v}(t), \ldots, N_D^{f-v}(t))^T$. Each video V_j gives one trial of $\mathbf{N}^{f-v}(t)$ with counting vector $(h_{j1}^{f-v}(t), h_{j2}^{f-v}(t), \ldots, h_{jD}^{f-v}(t))^T$, where $h_{jk}^{f-v}(t)$ is the value over the k-th bin of the histogram H_{jt}^{f-v}.

Similarly, D-dim multivariate point processes $\mathbf{N}^{f-a}(t)$, $\mathbf{N}^{b-v}(t)$, 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 [15] 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 [27], 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 [27] 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:

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

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 [37] 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):

$$\begin{aligned} S_{kl}(f) &= \frac{1}{2\pi MT}\sum\nolimits_{m=1}^{M}\tilde{N}_k(f,m)\tilde{N}_l(f,m)^*, \\ \tilde{N}_k(f,m) &= \sum\nolimits_{t_p=1}^{T} q_m(t_p)N_k(t_p)\exp(-2\pi i f t_p). \end{aligned} \tag{1}$$

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:

$$\begin{aligned} Y_1(t) &= \sum\nolimits_{p=1}^{\infty} a_p Y_1(t-p) + \sum\nolimits_{p=1}^{\infty} b_p Y_2(t-p) + \epsilon(t), \\ Y_2(t) &= \sum\nolimits_{p=1}^{\infty} c_p Y_2(t-p) + \sum\nolimits_{p=1}^{\infty} d_p Y_1(t-p) + \eta(t), \end{aligned}$$

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

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

We can compute the spectral matrix as [10]:

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

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 Eq. (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 [10]:

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

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

The Granger causality scores over all frequencies are then summed together to obtain a single time-domain causal influence, i.e., $C_{N_k \to N_l} = \sum_f G_{N_k \to N_l}(f)$, and $C_{N_l \to N_k} = \sum_f G_{N_l \to N_k}(f)$. In general, $C_{N_k \to N_l} \neq C_{N_l \to N_k}$, 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 $C_{N_k^{s_k} \to N_l^{s_l}}$ and $C_{N_l^{s_l} \to N_k^{s_k}}$ are summed together to generate the final similarity between these two codewords:

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

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:

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

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.

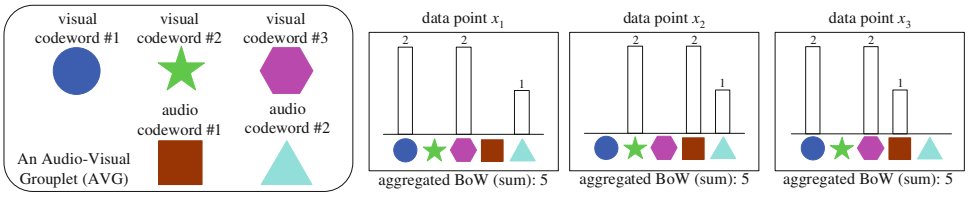


Fig. 5 An example of the aggregated BoW feature based on an AVG. In the example, assume that all codewords have equal weights, data points x_1, x_2, and x_3 have the same aggregated BoW features for the given AVG (value 5 by taking summation). However, data points x_1 and x_3 should be more similar to each other than data points x_1 and x_2. This is because x_1 and x_3 have the same feature values over visual codeword #1 and visual codeword #3, while x_1 and x_2 only have the same feature value over audio codeword #2

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 [28] 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.

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 Fig. 3, all of these correlations reveal useful audio-visual patterns for classifying concepts.

One intuitive way of using the AVGs for concept classification is to generate a feature value corresponding to each AVG for a given video. For instance, for the audio and/or visual codewords associated with an AVG, the values over the corresponding bins in the original visual-based and/or audio-based BoW features can be aggregated together (e.g., by taking summation or average) as the feature for the AVG. However, as illustrated in Fig. 5, such aggregated BoW features can be problematic and cannot fully utilize the advantage of the grouplet structure. In the next Sect. 7, we develop a distance metric learning algorithm to better use the AVGs for classifying concepts.

7 Grouplet-based distance metric learning

Assume that we have K AVGs G_k, $k = 1, \ldots, K$ in an audio-visual dictionary $\mathcal{D}$, where we temporarily omit upper indexes $(f-v, f-a)$, $(f-v, b-a)$, $(b-v, f-a)$, and $(b-v, b-a)$ w.o.l.g., since the grouplet-based distance metric learning algorithm will be applied to each dictionary individually. Let $D_k^G(x_i, x_j)$ denote the distance between data x_i and x_j computed based on the AVG G_k. The overall distance $D(x_i, x_j)$ between data x_i and x_j is given by:

$$D(x_i, x_j) = \sum_{k=1}^{K} v_k D_k^G(x_i, x_j). \tag{7}$$

The SVM classifiers with RBF-like kernels (Eq. 8) are found to provide state-of-the-art performances in several semantic concept classification tasks [19,34],

$$K(x_i, x_j) = \exp\left\{-\gamma D(x_i, x_j)\right\}. \tag{8}$$

For example, the chi-square RBF kernel usually performs well with histogram-like features [18,19], where distance $D(x_i, x_j)$ in Eq. (8) is the chi-square distance.

It is not trivial, however, to directly learn the optimal weights v_k $(k = 1, \ldots, K)$ in the SVM optimization setting, due to the exponential function in RBF-like kernels.

In this work, we formulate an iterative QP problem to learn optimal weights v_k $(k = 1, \ldots, K)$. The basic idea is to incorporate the LMNN setting for distance metric learning [40]. The rationale is that the role of large margin in LMNN is inspired by its role in SVMs, and LMNN should inherit various strengths of SVMs [33]. Therefore, although we do not directly optimize v_k $(k = 1, \ldots, K)$ in the SVM optimization setting, the final optimal weights can still provide reasonably good performance for SVM concept classifiers.

7.1 The LMNN formulation

Let $d_{\mathbf{M}}^2(x_i, x_j)$ denote the Mahalanobis distance metric between two data points x_i and x_j:

$$d_{\mathbf{M}}^2(x_i, x_j) = (x_i - x_j)^T \mathbf{M}(x_i - x_j), \tag{9}$$

where $\mathbf{M} \geq 0$ is a positive semi-definite matrix. LMNN learns an optimal $\mathbf{M}$ over a set of training data (x_i, y_i), $i = 1, \ldots, N$, where $y_i \in \{1, \ldots, c\}$ and c is the number of classes. For LMNN classification, the training process has two steps. First, n_k similarly labeled target neighbors are identified for each input training datum x_i. The target neighbors are selected by using prior knowledge or by simply computing n_k nearest (similarly labeled) neighbors using the Euclidean distance. Let $\eta_{ij} = 1$ (or 0) denote that x_j is a target neighbor of x_i (or not). In the second step, the

Mahalanobis distance metric is adapted so that these target neighbors are closer to x_i than all other differently labeled inputs. The Mahalanobis distance metric can be estimated by solving the following problem:

$$\min_{\mathbf{M}} \sum_{ij} \eta_{ij} \left[d^2_{\mathbf{M}}(x_i, x_j) + C \sum_l (1 - y_{il}) \epsilon_{ijl} \right],$$
$$\text{s.t. } d^2_{\mathbf{M}}(x_i, x_l) - d^2_{\mathbf{M}}(x_i, x_j) \geq 1 - \epsilon_{ijl}, \epsilon_{ijl} \geq 0, \mathbf{M} \geq 0.$$

$y_{il} \in \{0, 1\}$ indicates whether inputs x_i and x_l have the same class label. ϵ_{ijl} is the amount by which a differently labeled input x_l invades the "perimeter" around x_i defined by its target neighbor x_j.

7.2 Our approach

By defining $\mathbf{v} = [v_1, \ldots, v_K]^T$, $\mathbf{D}(x_i, x_j) = [D^G_1(x_i, x_j), \ldots, D^G_K(x_i, x_j)]^T$, we obtain the following problem:

$$\min_{\mathbf{v}} \left\{ \frac{||\mathbf{v}||_2^2}{2} + C_0 \sum_{ij} \eta_{ij} \mathbf{v}^T \mathbf{D}(x_i, x_j) + C \sum_{ijl} \eta_{ij} (1 - y_{il}) \epsilon_{ijl} \right\},$$
$$\text{s.t. } \mathbf{v}^T \mathbf{D}(x_i, x_l) - \mathbf{v}^T \mathbf{D}(x_i, x_j) \geq 1 - \epsilon_{ijl}, \epsilon_{ijl} \geq 0, v_k \geq 0.$$

$||\mathbf{v}||_2^2$ is the L_2 regularization that controls the complexity of $\mathbf{v}$. By introducing Lagrangian multipliers $\mu_{ijl} \geq 0$, $\gamma_{ijl} \geq 0$, and$\sigma_k \geq 0$, we have:

$$\min_{\mathbf{v}} \left\{ \frac{||\mathbf{v}||_2^2}{2} + C_0 \sum_{ij} \eta_{ij} \mathbf{v}^T \mathbf{D}(x_i, x_j) \right.$$
$$- \sum_{ijl} \mu_{ijl} \eta_{ij} \left[\mathbf{v}^T \mathbf{D}(x_i, x_l) - \mathbf{v}^T \mathbf{D}(x_i, x_j) - 1 + \epsilon_{ijl} \right]$$
$$\left. - \sum_{ijl} \gamma_{ijl} \eta_{ij} \epsilon_{ijl} - \sum_k \sigma_k v_k + C \sum_{ijl} \eta_{ij} (1 - y_{il}) \epsilon_{ijl} \right\}. \quad (10)$$

Next, by taking derivative against ϵ_{ijl} we obtain:

$$C \eta_{ij} (1 - y_{il}) - \mu_{ijl} \eta_{ij} - \gamma_{ijl} \eta_{ij} = 0. \quad (11)$$

That is, for any pair of x_i and its target neighbor x_j, since we only consider x_l with $y_{il} = 0$, $0 \leq \mu_{ijl} \leq C$. Based on Eq. (11), Eq. (10) turns to:

$$\min_{\mathbf{v}} \left\{ \frac{1}{2} ||\mathbf{v}||_2^2 + C_0 \sum_{ij} \eta_{ij} \mathbf{v}^T \mathbf{D}(x_i, x_j) \right.$$
$$\left. - \sum_{ijl} \mu_{ijl} \eta_{ij} \left[\mathbf{v}^T \mathbf{D}(x_i, x_l) - \mathbf{v}^T \mathbf{D}(x_i, x_j) - 1 \right] - \mathbf{v}^T \sigma \right\}, \quad (12)$$

where $\sigma = [\sigma_1, \ldots, \sigma_K]^T$. Then by taking derivative against $\mathbf{v}$ we get:

$$\mathbf{v} = \sum_{ijl} \mu_{ijl} \eta_{ij} \left[\mathbf{D}(x_i, x_l) - \mathbf{D}(x_i, x_j) \right] + \sigma - C_0 \sum_{ij} \eta_{ij} \mathbf{D}(x_i, x_j). \quad (13)$$

Define set $\mathcal{P}$ as the set of indexes i, j, l that satisfy the conditions of $\eta_{ij} = 1$, $y_{il} = 0$, and that x_l invades the "perimeter" around the input x_i defined by its target neighbor x_j, i.e., $0 \leq D(x_i, x_l) - D(x_i, x_j) \leq 1$. Define set $\mathcal{Q}$ as the set of indexes i, j that satisfy $\eta_{ij} = 1$. Next, we can use μ_p, $p \in \mathcal{P}$ to replace the original notation μ_{ijl}, use $\mathbf{D}^p_{\mathcal{P}}$, $p \in \mathcal{P}$ to replace the corresponding $\mathbf{D}(x_i, x_l) - \mathbf{D}(x_i, x_j)$, and use $\mathbf{D}^q_{\mathcal{Q}}$, $q \in \mathcal{Q}$ to replace the corresponding $\mathbf{D}(x_i, x_j)$. Define $\mathbf{u} = [\mu_1, \ldots, \mu_{|\mathcal{P}|}]^T$, $|\mathcal{P}| \times K$ matrix $\mathbf{D}_{\mathcal{P}} = \left(\mathbf{D}^1_{\mathcal{P}}, \ldots, \mathbf{D}^{|\mathcal{P}|}_{\mathcal{P}} \right)^T$, and $|\mathcal{Q}| \times K$ matrix $\mathbf{D}_{\mathcal{Q}} = \left(\mathbf{D}^1_{\mathcal{Q}}, \ldots, \mathbf{D}^{|\mathcal{Q}|}_{\mathcal{Q}} \right)^T$. Through some derivation, we obtain the dual of Eq. (12) as follows:

$$\max_{\sigma, \mathbf{u}} \left\{ -\frac{1}{2} \mathbf{u}^T \mathbf{D}_{\mathcal{P}} \mathbf{D}^T_{\mathcal{P}} \mathbf{u} + C_0 \mathbf{u}^T \mathbf{D}_{\mathcal{P}} \mathbf{D}^T_{\mathcal{Q}} \mathbf{1}_{\mathcal{Q}} + \mathbf{u}^T \mathbf{1}_{\mathcal{P}} \right.$$
$$\left. - \frac{1}{2} \sigma^T \sigma - \mathbf{u}^T \mathbf{D}_{\mathcal{P}} \sigma + C_0 \sigma^T \mathbf{D}^T_{\mathcal{Q}} \mathbf{1}_{\mathcal{Q}} \right\}, \quad (14)$$

where $\mathbf{1}_{\mathcal{Q}}$ ($\mathbf{1}_{\mathcal{P}}$) is a $|\mathcal{Q}|$-dim ($|\mathcal{P}|$-dim) vector whose elements are all ones.

When σ is fixed, Eq. (14) can be further rewritten to the following QP problem:

$$\max_{\mathbf{u}} \left\{ \frac{1}{2} \mathbf{u}^T \mathbf{D}_{\mathcal{P}} \mathbf{D}^T_{\mathcal{P}} \mathbf{u} + \mathbf{u}^T \left(C_0 \mathbf{D}_{\mathcal{P}} \mathbf{D}^T_{\mathcal{Q}} \mathbf{1}_{\mathcal{Q}} + \mathbf{1}_{\mathcal{P}} - \mathbf{D}_{\mathcal{P}} \sigma \right) \right\},$$
$$\text{s.t. } \forall p \in \mathcal{P}, 0 \leq \mu_p \leq C. \quad (15)$$

On the other hand, when $\mathbf{u}$ is fixed, Eq. (14) turns into the following QP problem:

$$\max_{\sigma} \left\{ -\frac{1}{2} \sigma^T \sigma + \sigma^T \left(C_0 \mathbf{D}^T_{\mathcal{Q}} \mathbf{1}_{\mathcal{Q}} - \mathbf{D}^T_{\mathcal{P}} \mathbf{u} \right) \right\},$$
$$\text{s.t. } \forall k = 1, \ldots, K, \sigma_k \geq 0. \quad (16)$$

Therefore, we can iteratively solve the QP problems of Eqs. (15) and (16) and obtain the desired weights $\mathbf{v}$ through Eq. (13).

For each of the QP problems, since we have positive definite Q (or positive semi-definite Q that can be made positive definite by using practical tricks), it can be solved efficiently in polynomial time.

7.3 Grouplet-based kernels

One of the most intuitive kernels that incorporates the AVG information is the grouplet-based chi-square RBF kernel. That is, each $D^G_k(x_i, x_j)$ is a chi-square distance:

$$D^G_k(x_i, x_j) = \sum_{w_m \in G_k} \frac{\left[f_{w_m}(x_i) - f_{w_m}(x_j) \right]^2}{\frac{1}{2} \left[f_{w_m}(x_i) + f_{w_m}(x_j) \right]}, \quad (17)$$

where $f_{w_m}(x_i)$ is the feature of x_i corresponding to the codeword w_m in AVG G_k. When $v_k = 1$, $k = 1, \ldots, K$, Eq. (17) will give the standard chi-square RBF kernel.

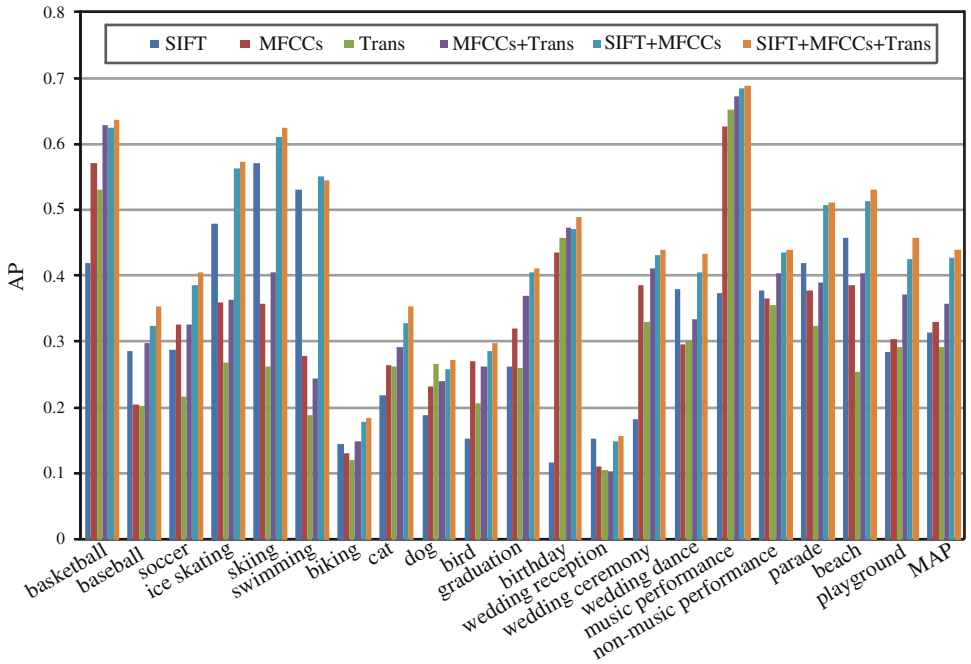


Fig. 6 Comparison of various BoW representations as well as their early-fusion combinations

From another perspective, we can treat each AVG as a *phrase*, which consists of the orderless codewords associated with that AVG. Analogous to measuring the similarity between two text segments, we should take into account the word specificity [26] in measuring the similarity between data points. One popular way of computing the word specificity is to use the inverse document frequency (idf). Therefore, we use the following metric to compute $D_k^G(x_i, x_j)$:

$$\frac{1}{\sum_{w_m \in G_k} \text{idf}(w_m)} \sum_{w_m \in G_k} \text{idf}(w_m) \frac{\left[f_{w_m}(x_i) - f_{w_m}(x_j)\right]^2}{\frac{1}{2}\left[f_{w_m}(x_i) + f_{w_m}(x_j)\right]}. \tag{18}$$

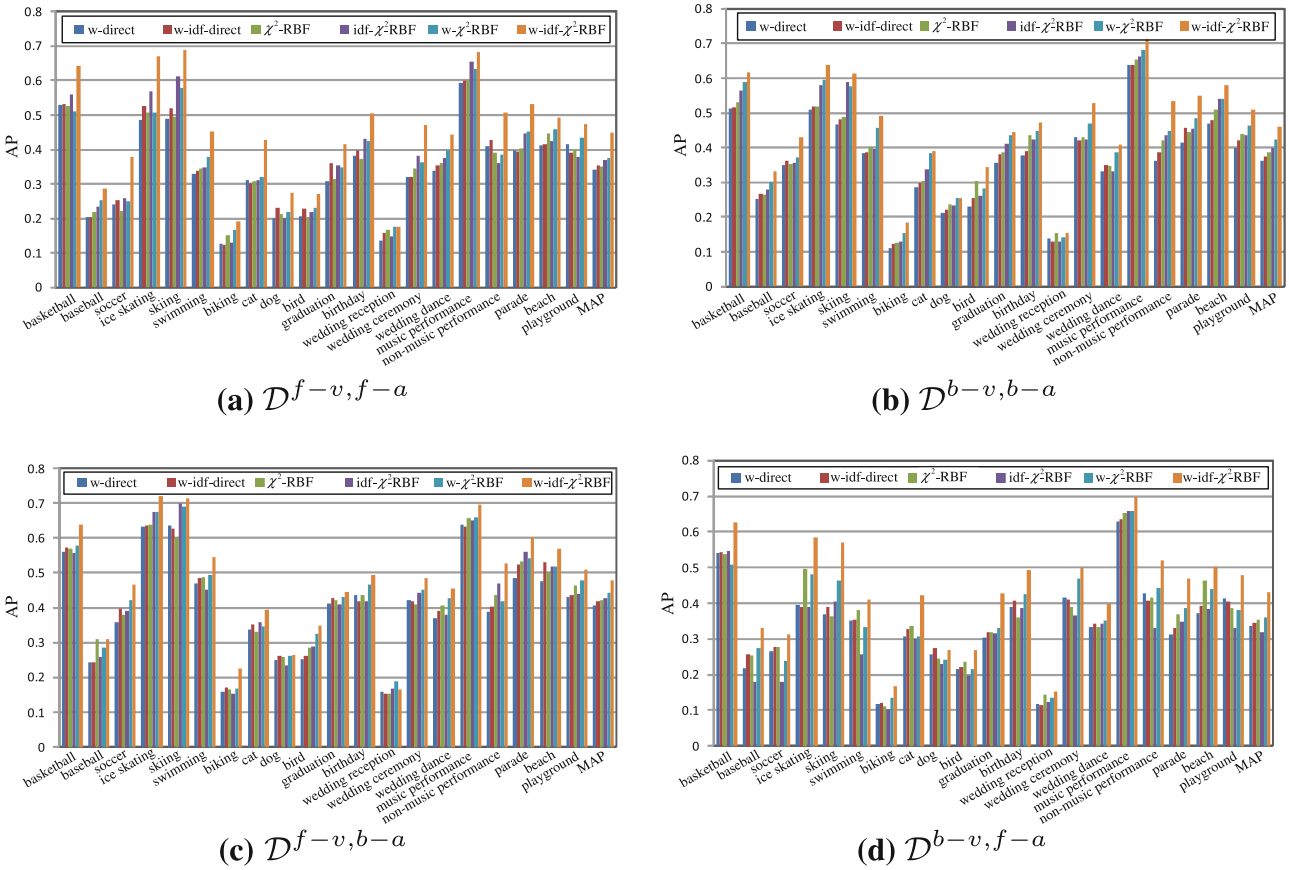
$\text{idf}(w_m)$ is computed as the total number of occurrences of all codewords in the training corpus divided by the total number of occurrences of w_m in the training corpus. Using either the chi-square distance Eq. (17) or the idf-weighted chi-square distance Eq. (18), respectively, the distance metric learning method developed in the previous Sect. 7.2 can be applied to find the optimal metric and compute the optimal kernels for concept classification.

Finally, as described in Sect. 6, we have four types of audio-visual dictionaries by studying four types of audio-visual temporal correlations. The distance metric learning algorithm described in Sect. 7.2 can be applied to each type of dictionary individually, and four types of optimal kernels can be computed. After that, the Multiple Kernel Learning technique [36] is adopted to combine the four types of kernels for final concept detection.

8 Experiments

We evaluate our algorithm over the large-scale CCV set [19], containing 9317 consumer videos from YouTube. The 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 [22,34]. 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 [19]. Our experiments take similar settings as [19], i.e., we use the same training (4659 videos) and test (4658 videos) sets, and one-versus-all SVM classifiers. The performance is measured by Average Precision (AP, the area under uninterpolated PR curve) and Mean AP (MAP, averaged AP across concepts).

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 Fig. 6. These BoW representations are generated using the same method as [19]. 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



(a) $\mathcal{D}^{f-v,f-a}$ **(b)** $\mathcal{D}^{b-v,b-a}$

(c) $\mathcal{D}^{f-v,b-a}$ **(d)** $\mathcal{D}^{b-v,f-a}$

Fig. 7 Performances comparison of different approaches using individual types of audio-visual dictionaries

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 [19]. 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 1×1 spatial layout, while several types of keypoints and spatial layouts are used in [19]. Actually, our "MFCCs" performs similarly to that in [19], due to the similar settings for feature extraction and vocabulary construction.

Next, we show the classification performance of using different types of individual 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 $200 \sim 300$ AVGs on average. The results are shown in Fig. 7a–d. The goal is to demonstrate the usefulness of the proposed AVG-based distance metric learning algorithm. Here we compare 6 different approaches: the standard chi-square RBF kernel ("χ^2-RBF"); the "w-direct" method that uses distance metric learning to directly combine χ^2 distances computed over individual audio and visual vocabularies; the chi-square RBF kernel that uses the idf information ("idf-χ^2-RBF"); the "w-idf-direct" method that uses distance metric learning to directly combine idf-weighted χ^2 distances computed over individual audio and visual vocabularies; the weighted chi-square RBF kernel with distance metric learning that uses the AVGs ("w-χ^2-RBF"); and the weighted chi-square RBF kernel with distance metric learning that uses both the idf information and the AVGs ("w-idf-χ^2-RBF"). In other words, "w-direct", "w-idf-direct", "χ^2-RBF" and "idf-χ^2-RBF" do not use any AVG information. From the figures we can see that by finding appropriate weights of AVGs through our distance metric learning, we can consistently improve the detection performance. For example, for all four types of AVGs, "w-χ^2-RBF" works better than "χ^2-RBF" on average, and "w-idf-χ^2-RBF" outperforms "idf-χ^2-RBF." Also, the advantages of "w-idf-χ^2-RBF" are quite apparent, i.e., it performs the most efficiently over almost every concept across all types of AVGs. In comparison, without generating the AVGs, by directly applying distance metric learning to combine individual audio and visual vocabularies, "w-direct" and "w-idf-direct" cannot bring any overall improvements. One possible reason is due to the large amount of parameters to learn for distance metric learning in such cases, e.g., 4000 for each type of vocabulary, one corresponding to each

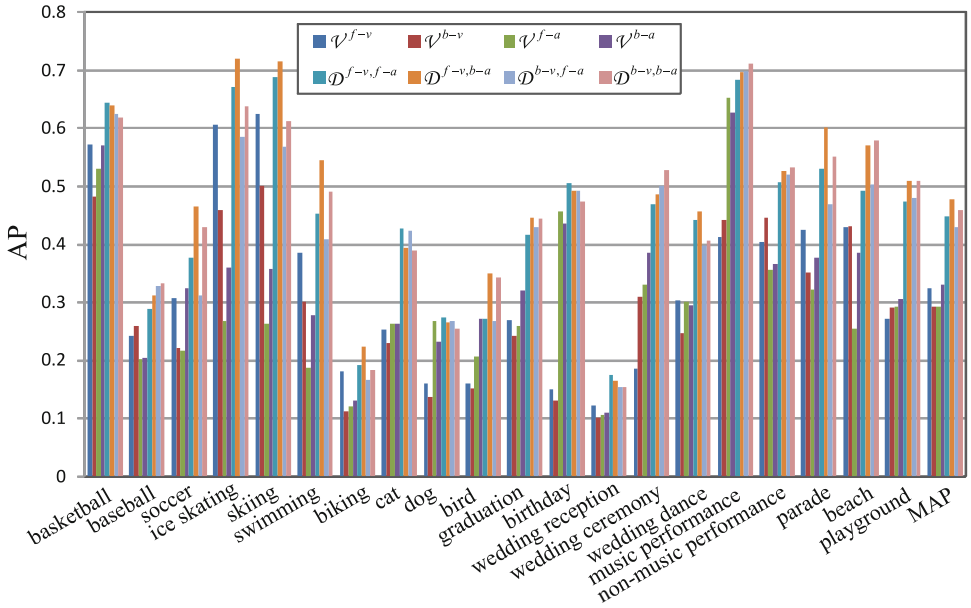


Fig. 8 Comparison of individual foreground/background audio/visual vocabularies and audio-visual dictionaries

feature dimension. This problem is effectively alleviated by incorporating the AVG representation, where the amount of parameters to learn is largely reduced.

Then, 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 $\mathcal{V}^{b-a}$) via the BoW representation, as well as using various types of individual audio-visual dictionaries via "w-idf-χ^2-RBF" kernel. The results are given in Fig. 8. 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 [5,19]. 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 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 temporal audio-visual interactions, 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 $\mathcal{V}^{f-a}$, on a relative basis, by roughly 40 and 50 %, respectively, and the MAP of $\mathcal{D}^{b-v,b-a}$ outperforms those of $\mathcal{V}^{b-v}$ and $\mathcal{V}^{b-a}$ by roughly 50 and 40 %, respectively.

Finally, the four types of audio-visual dictionaries are combined together to train concept classifiers so that the advantages of all dictionaries in classifying different concepts can be exploited. Figure 9 shows the final performance, where multiple kernel learning is applied to find the optimal weights to combine kernels computed over individual audio-visual dictionaries. Here we compare our "MKL-w-idf-χ^2-RBF" approach with three other alternatives: the early fusion of the BoW representations from multiple types of features ("SIFT+MFCCs+Trans"), which is considered the state-of-the-art in the literature; the "MKL-Vocabulary" method where multiple kernel learning is used to combine standard χ^2-RBF kernels computed over the four types of individual audio and visual foreground and background vocabularies; and the "MKL-Dictionary" method where multiple kernel learning is used to combine χ^2-RBF kernels computed over individual audio-visual dictionaries based on the AVG-based features generated by aggregating BoW bins, as described in Fig. 5. From the figure we can see that our "MKL-w-idf-χ^2-RBF" can consistently and significantly outperform other alternatives over all concepts. Compared with "SIFT+MFCCs+Trans," "w-idf-χ^2-RBF" improves the overall MAP by more than 20 %, and significant AP gains (more than 20 %) are obtained over 12 concepts, e.g., roughly 40 % gain over "basketball," 60 % gain over "biking," 40 % gain over "wedding reception," 40 % gain over "wedding ceremony," and 40 % gain over "non-music performance." Compared with the "MKL-Dictionary" that uses AVGs in the naive way and the "MKL-Vocabulary" that does not use the AVG information, we improve the AP of every concept by more than 5 %, and over 15 concepts, the improves are more than 10 %. The results demonstrate the effectiveness of extracting useful AVGs to represent general videos and

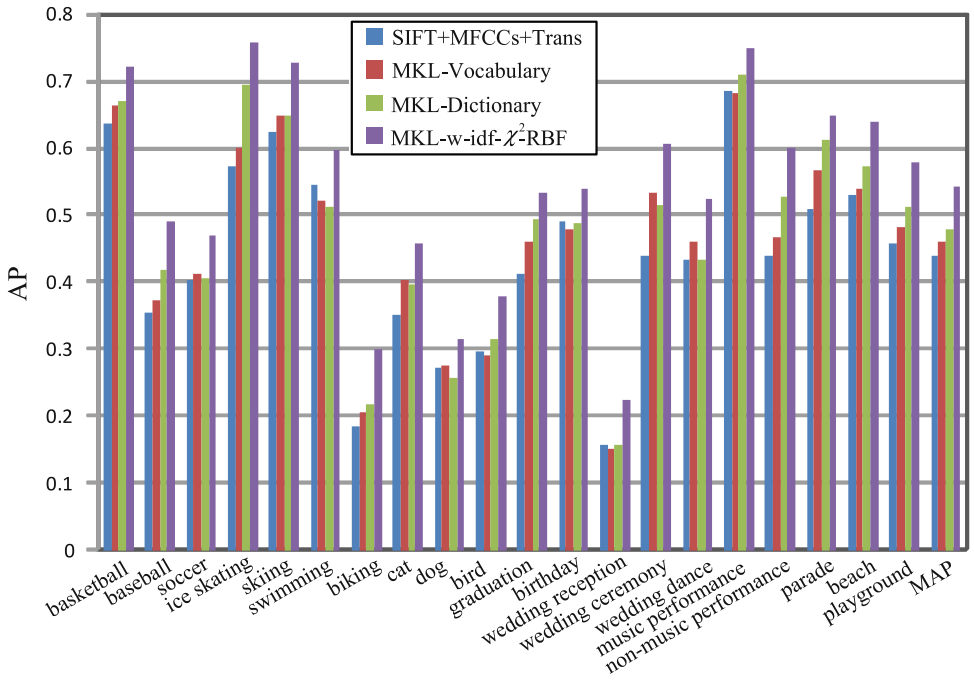


Fig. 9 Combining different types of audio-visual dictionaries

using AVG-based distance metric learning for concept classification.

The training process of generating AVGs is relatively expensive in computation, where the most time consuming part lies in the processes of conducting SIFT tracking and computing causal matrices. However, once the AVGs are obtained, the classification process of using such AVGs can be reasonably fast. Specifically, the complexity of generating BoW vectors as well as concept classification are similar to standard acts in the field, and we can reduce the sample frequency in conducting SIFT tracking to alleviate the computational overhead. In addition, the number of AVGs (hundreds) is usually much smaller than the original number of codewords (thousands) in the audio and visual vocabularies, and the final SVM classification is faster than traditional BoW approaches. On average, the classification of test videos can be real-time, i.e., it takes about 1 min to classify 20 concepts over a 1-min long video, using a dual-core machine with 8G ram.

9 Conclusion

An AVG representation is proposed 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. We conduct coarse-level foreground/background separation in both visual and audio channels, and extract four types of AVGs based on 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. To use the AVGs for effective concept classification, a distance metric learning algorithm was further developed. Based on the LMNN setting, the algorithm optimizes an iterative QP problem to find the appropriate weights of combining individual grouplet-based distances for optimal classification. 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. Also, the L_2 regularization of weights $\mathbf{v}$ is used in our distance metric learning algorithm to prevent sparse solutions, due to the relatively small number of AVGs in our experiments. For tasks with a large number of AVGs, L_1-norm that encourages sparsity may be a better choice. In addition, the spatial relations of visual SIFT tracks can be incorporated to further help classification. The spatial-temporal audio-visual correlations can be explored in the future, e.g., by constructing spatially-correlated visual signatures first and then correlating such visual signatures with audio codewords.

Acknowledgments We would like to thank the authors of [6] and [27] for sharing their code with us, and for Shih-Fu Chang for many useful discussions.

References

1. Aach T, Kaup A (1995) Bayesian algorithms for adaptive change detection in image sequences using Markov random fields. Signal Process: Image Commun 7:147–160
2. Akutsu A, Tonomura Y (1994) Video tomography: an efficient method for camerawork extraction and motion analysis. In: ACM multimedia, pp 349–356
3. Barzelay Z, Schechner Y (2007) Harmony in motion. In: IEEE CVPR, pp 1–8
4. Beal MJ, Jojic N, Attias H (2003) A graphical model for audiovisual object tracking. IEEE PAMI 25(7):828–836
5. Chang S et al (2007) Large-scale multimodal semantic concept detection for consumer video. In: ACM MIR, pp 255–264
6. Cotton C, Ellis D, Loui A (2011) Soundtrack classification by transient events. In: IEEE ICASSP, Czech Republic
7. Cristani M, Manuele B, Vittorio M (2007) Audio-visual event recognition in surveillance video sequences. IEEE Trans Multimedia 9(2):257–267
8. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE CVPR, pp 886–893
9. Davis J et al (2007) Information-theoretic metric learning. In: ICML, pp 209–216
10. Ding M, Chen Y, Bressler SL (2006) Granger causality: basic theory and applications to neuroscience. In: Schelter S et al (eds) Handbook of time series analysis. Wiley, Wienheim
11. Divvala S et al (2009) An empirical study of context in object detection. In: IEEE CVPR, Miami
12. Elhabian SY, El-Sayed KM (2008) Moving object detection in spatial domain using background removal techniques: state-of-art. Recent Patents Comput Sci 1(1):32–54
13. Enqvist O, Josephson K, Kahl F (2009) Optimal correspondences from pairwise constraints. In: IEEE ICCV, Kyoto
14. Globerson A, Roweis S (2006) Metric learning by collapsing classes. In: NIPS, pp 451–458
15. Granger C (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37(3):424–438
16. Iwano K et al (2007) Audio-visual speech recognition using lip information extracted from side-face images. EURASIP J ASMP 2007(1):4–12
17. Jiang W et al (2010) Audio-visual atoms for generic video concept classification. ACM TOMCCAP 6:1–19
18. Jiang Y, Ngo CW, Yang J (2007) Towards optimal bag-of-features for object categorization and semantic video retrieval. In: ACM CIVR, pp 494–501
19. Jiang Y et al (2011) Consumer video understanding: a benchmark database and an evaluation of human and machine performance. ACM ICMR, Trento
20. Joly P, Kim HK (1996) Efficient automatic analysis of camera work and microsegmentation of video using spatiotemporal images. Signal Process: Image Commun 8(4):295–307
21. Ke Y, Sukthankar R, Hebert M (2007) Event detection in crowded videos. IEEE ICCV, Brazil
22. Laptev I et al (2008) Learning realistic human actions from movies. IEEE CVPR, Alaska
23. Lin WH, Hauptmann A (2002) News video classification using svm-based multimodal classifiers and combination strategies. In: Proc ACM multimedia, pp 323–326
24. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. IJCV 60(2):91–110
25. Marszalek M, Laptev I, Schmid C (2009) Actions in context. In: IEEE CVPR, Miami
26. Mihalcea R, Corley C, Strapparava C (2006) Corpus-based and knowledge-based measures of text semantic similarity. In: National conference on artificial intelligence, pp 775–780
27. Nedungadi A et al (2009) Analyzing multiple spike trains with nonparametric granger causality. J Comput Neurosci 27(1):55–64
28. Ng A, Jordan M, Weiss Y (2001) On spectral clustering: analysis and an algorithm. NIPS
29. Pham DT, Cardoso JF (2001) Blind separation of instantaneous mixtures of non stationary sources. IEEE Trans Signal Process 49(9):1837–1848
30. Prabhakar K et al (2010) Temporal causality for the analysis of visual events. In: IEEE CVPR, San Francisco
31. Roweis ST (2001) One microphone source separation. NIPS
32. Sargin M et al (2009) Audiovisual celebrity recognition in unconstrained web videos. In: IEEE ICASSP, Taipei
33. Schölkopf B, Smola AJ (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT, Cambridge
34. Smeaton AF, Over P, Kraaij W (2006) Evaluation campaigns and TRECVid. In: ACM MIR, pp 321–330
35. Stauffer C, Grimson E (2000) Learning patterns of activity using realtime tracking. IEEE PAMI 22(8):747–757
36. Varma M, Babu BR (2009) More generality in efficient multiple kernel learning. In: ICML, pp 1065–1072
37. Walden A (2000) A unified view of multitaper multivariate spectral estimation. Biometrika 87(4):767–788
38. Wang B, Plumbley MD (2006) Investigating single-channel audio source separation methods based on non-negative matrix factorization. In: ICArn, pp 17–20
39. Wang W et al (2005) Video assisted speech source separation. In: IEEE ICASSP, pp 425–428
40. Weinberger K, Saul L (2009) Distance metric learning for large margin nearest neighbor classification. JMLR 10(12):207–244
41. Wu L et al (2009) Scale-invariant visual language modeling for object categorization. IEEE TMM 11(2):286–294
42. Wu L et al (2010) Semantics-preserving bag-of-words models and applications. IEEE TIP 19(7):1908–1920
43. Wu Y et al (2004) Multimodal information fusion for video concept detection. IEEE ICIP, pp 2391–2394
44. Yang J et al (2007) Evaluating bag-of-visual-words representations in scene classification. ACM MIR, pp 197–206
45. Yao B, Fei-Fei L (2010) Grouplet: a structured image representation for recognizing human and object interactions. IEEE CVPR, San Francisco