Hindawi Publishing Corporation
Advances in Multimedia
Volume 2015, Article ID 698316, 10 pages
http://dx.doi.org/10.1155/2015/698316

Research Article

High-Level Codewords Based on Granger Causality for Video Event Detection

Shao-nian Huang,[1,2] Dong-jun Huang,[1] and Mansoor Ahmed Khuhro[1]

[1]*School of Information Science and Engineering, Central South University, Changsha 410083, China*
[2]*School of Computer and Information Engineering, Hunan University of Commerce, Changsha 410205, China*

Correspondence should be addressed to Shao-nian Huang; hsn@hunnu.edu.cn

Received 18 January 2015; Revised 19 May 2015; Accepted 7 June 2015

Academic Editor: Luigi Atzori

Video event detection is a challenging problem in many applications, such as video surveillance and video content analysis. In this paper, we propose a new framework to perceive high-level codewords by analyzing temporal relationship between different channels of video features. The low-level vocabulary words are firstly generated after different audio and visual feature extraction. A weighted undirected graph is constructed by exploring the Granger Causality between low-level words. Then, a greedy agglomerative graph-partitioning method is used to discover low-level word groups which have similar temporal pattern. The high-level codebooks representation is obtained by quantification of low-level words groups. Finally, multiple kernel learning, combined with our high-level codewords, is used to detect the video event. Extensive experimental results show that the proposed method achieves preferable results in video event detection.

1. Introduction

With the increasing popularity of digital cameras and mobile phones, more and more consumer-generated web videos recording real-life events are widely available on Internet. For example, more than 100 hours of videos are uploaded to YouTube every minute [1]. Consequently, how to effectively manage and retrieve the unconstrained consumer videos is becoming an urgent problem. In particular, video event recognition is receiving increasing attention in the field of computer vision [2]. However, it is an extremely difficult task due to the different video content and the variable conditions in lighting, camera motion, and occlusions. Figure 1 shows some representative frames from events "bird" defined in Columbia Consumer Video (CCV) Database [3]. We can see that the contents of these six videos are dramatically different, although they are all belonging to the same type of event.

The majority of existing event-recognition methods classified video mainly based on visual information. In general, various visual features of key frames were extracted for event classification [4, 5]. Some other event detection methods used high-level visual feature representation which modeled the relationship between low-level visual features and semantic concepts [6, 7]. But in fact, besides visual features, audio information of the same video also provides important cue for event recognition [8, 9].

To better describe the underlying causality in videos, in this work, we propose a high-level codebooks representation utilizing the Granger's Causality [10] between different channels of features. First, the low-level visual feature and audio features are extracted, which are clustered to form visual bag-of-words (BoW) and audio BoW, respectively. To model the temporal causality between the two channels of information, the vocabulary representation of video sequence is viewed as the instantaneous of multivariate point process. By analyzing the Granger Causality between low-level audio and visual words, an undirected weighted graph is constructed to model the temporal causality of the videos. After that, we split the graph into low-level word groups which indicate the temporal patterns in videos. Finally the high-level codebooks are generated by quantifying low-level word groups, and then video event is detected based on multiple kernel learning framework (MKL) [11]. We evaluate our method on public datasets and perform the comparison with a number of other

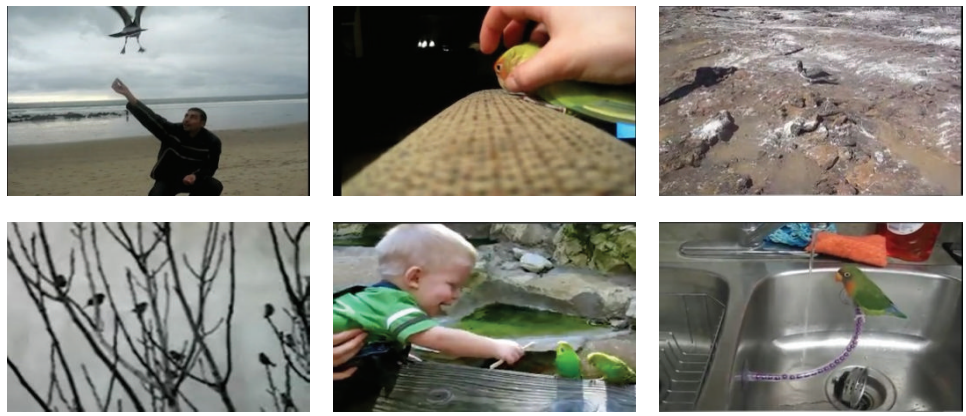

Figure 1: Example video frames of event "bird" defined in Columbia Consumer Video Database.

state-of-the-art methods. The experimental results illustrate that the method proposed in this paper can achieve preferable results in consumer videos. The general flowchart of our proposed method is shown in Figure 2.

In summary, the main contributions of this paper are (1) a proposed framework to perceive high-level codewords by taking the temporal causality between low-level features into account; (2) construction of a temporal relationship graph to extract high-level codewords; (3) utilization of the multiple kernel learning framework to detect video events.

The rest of the paper is organized as follows. We first review related works, especially the popular feature fusion method used in video event detection in Section 2. We continue with extraction of low-level video feature in Section 3. In Section 4, we propose a high-level codewords framework for event detection based on the temporal causality. Our experimental results on public datasets are provided in Section 5. The paper ends with conclusions and prospects for future work in Section 6.

2. Related Work

Multiple feature fusion for multimedia analysis has been extensively studied. Compared to using only single feature, multifeature fusion has been proven to enhance the performance for multimedia content analysis. General speaking, early fusion and late fusion are the two popular ways for feature combination [12]. Early fusion concatenates features from different modalities into a single vector, while late fusion combines the results of different classifier to obtain the final classification score by a certain principle. However, the question on how to construct suitable joint feature and classifier combination still remains an open issue.

In the fields of machine learning, many researchers have been devoted to develop multiple view of learning algorithms to achieve multiple feature fusion. In [13], a multiple feature fusion algorithm is proposed by learning a generalized subspace in which canonical correlation between low-level features is measured. Oh et al. designed a multimedia event detection framework based on Latent SVM model which can learn high-level concepts [14]. Multistage feature strategy has been exploited by Natarajan et al. for complex event detection, such as multiple kernel learning, score level fusion, and weighted average fusion [15]. However, majority of these methods may need a large amount of label training data, but the real-world videos often lack exact labels, especially in consumer video. The semisupervised learning method has been proven to efficiently use unlabeled data to infer an accurate classifier [16–18]. In [16], Yang et al. designed a hierarchical regression model to learning classifier which can utilize unlabeled data to represent multiple features. Recently, Ma et al. proposed a semisupervised learning framework with little-labeled training data by integrating multifeature learning and the Riemannian metric [17]. In [18], Xu et al. designed a cross-feature learning model for complex event detection based on the multilevel relevance learning of related exemplars.

Some other works concentrated on the use of audio-visual cue for tracking and recognition [19, 20]. Derbas and Quénot proposed an audio-visual feature representation to detect violent scenes in movies [19]. Ionescu et al. designed a content descriptor which includes audio and color content for video categorization [20]. However, the empirical results of these methods are subject to many qualifications, such as the category of the video and the environment of the video.

More recently, Jhuo et al. proposed an audio-visual bimodal representation for video event detection [21]. The audio-visual descriptors were firstly extracted to a construct bipartite graph discovering the joint probability of audio words and visual words. Bimodal words were then obtained by graph partition. Different from the above methods based on statistical relation, Prabhakar et al. firstly produced space-time dictionary by temporal causality for visual event analysis [22]. As an extension of this work, Jiang and Loui introduced an audio-visual grouplets representation method which uses the temporal audio-visual relation [23]. The author constructed four types of grouplets between the combination of foreground and background information. Despite the close relationship with our work, the above method requires visual foreground/background separation and audio background/foreground extraction, which remain extremely difficult and time consuming in consumer videos. In this paper, the proposed method is suitable for general Internet video and avoids region segmentation.

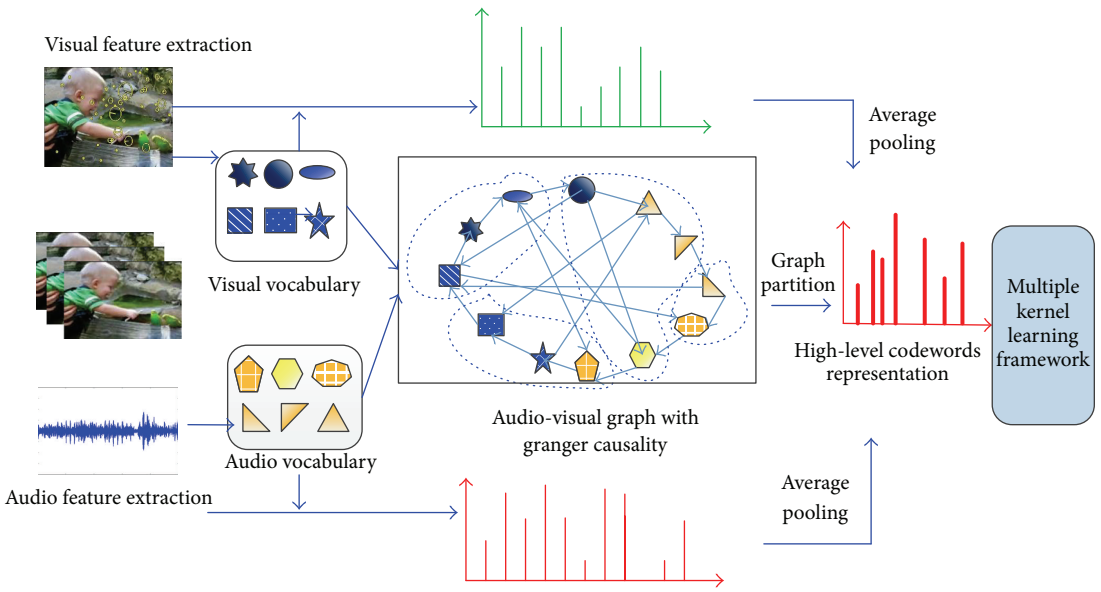

Figure 2: Flowchart of the proposed method.

3. Low-Level Vocabulary Representation

BoW approach is a popular feature representation method which had been proven to be surprisingly effective in video analysis [3]. In this paper, two types of low-level features are extracted from training videos and then generate two types of BoW of videos. We used the following low-level descriptors in our work.

SIFT. Scale-invariant feature transform (SIFT) has been widely used in many researches of video content analysis, such as object recognition and video concept detection [24], since it is invariant to image scale, rotation, and changing viewpoints. In this paper, the difference of Gaussians operator was adopted to find local keypoint in the frames. Then, a 128-dimensional feature descriptor at each point was formed to capture the local gradients. In order to reduce of computation cost, we extracted features from sampled frames with a sample rate of 3 frames per second.

STIP. As an important cue for video content analysis, the popular spatial-temporal interest point (STIP) extracts the local space-time structure where the image values have significant local variations in both space and time [25]. In this paper, the Harris 3D detector was adopted to locate space-time volume. Each volume was subdivided into a (n_x, n_y, n_t) grids of cuboids, and then 4 bins histograms of gradients (HOG) and 5 bins histograms of optical flow (HOF) were computed from the grids. The parameters are set as same as paper [25], such as $n_x = n_y = 3, n_t = 2$. Finally, we directly concatenated the HOG and HOF feature into a 162-dimensional vector which represents the local motion.

DTF. Dense trajectories feature (DTF) has been shown to be among the best visual feature in the application of video analysis [26]. Following the set in [26], we extracted the dense trajectories by the sampled feature points on a dense grid, and the trajectory descriptors were obtained by $N * N * L$ space-time volume around the trajectory. Finally we extracted 96-dimensional HOG feature of the trajectory.

MFCC. Acoustic features have been found to be very useful for various recognition systems. Among different acoustic features, mel-frequency cepstral coefficients (MFCC) [27], which collectively represents short-term power spectrum of sound based on a linear cosine transform, is one of the most prevalent choice for audio recognition. For each video, we extracted 36-dimensional MFCCs feature over 20 ms window size with 10 ms overlap.

Four low-level codebooks were generated by clustering the above features, respectively. For each video clip, the four features are quantified to form four BoW histogram representations. In order to discuss the temporal causality between low-level features, we directly concatenated different visual features to form visual BoW which provides the visual information of the video, while the MFCC BoW represents the audio information. They were used to extract high-level codewords as discussed in the next section.

4. High-Level Codewords Representation

In this section the high-level codewords representation based on Granger Causality is explained in detail. We first viewed each word in the video as a point process and analyzed the Granger Causality between low-level codewords. Then we constructed a weighted undirected graph based on the temporal relationship to extract high-level codewords. At the end, we used multiple kernel learning framework to detect video event.

4.1. Temporal Causality between Low-Level Codewords. Audio information is an important cue for video event

detection. The emergence of some type of visual objects always accompanies some kinds of background sound. For example, the audio background of a basketball match is often the ball bouncing sounds, and the appearance of a dog in video often follows barking. Therefore, we need to first analyze the temporal causality between visual information and audio information and then detect the video event based on the audio-visual relevance.

Prabhakar et al. were the first to propose a method to describe the temporal causality between the visual words in videos by viewing the words sequence as multivariate point process [22]. Here we use $w^a = \{w_1^a, w_2^a, \ldots, w_{N_a}^a\}$ and $w^v = \{w_1^v, w_2^v, \ldots, w_{N_v}^v\}$ to represent the sets of audio and visual vocabulary, respectively, where N_a and N_v denote the number of audio and visual words. In order to investigate the cooccurrence of w^a and w^v, we compute the probability of each word w_i^a and w_i^v within each video frame. Firstly, the amount of emergence of word w_i^a in the interval $(0, t]$ is defined as

$$dN_i^a(t) = N_i^a(t + dt) - N_i^a(t), \tag{1}$$

where dt denotes the time resolution. The mean intensity of the process $N_i^a(t)$ is defined as $E\{N_i^a(t)\} = \lambda_i^a dt$. Then we consider the zero-mean process $N_i^a(t) - \lambda_i^a \cdot dt$ and rename that process $N_i^a(t)$. Therefore, all N_a visual words create a N_a-dimensional multivariate point process $N^a(t) = \{N_1^a(t), N_2^a(t), \ldots, N_{N_a}^a(t)\}$. Similarly, N_v-dimensional multivariate point process $N^v(t) = \{N_1^v(t), N_2^v(t), \ldots, N_{N_v}^v(t)\}$ can be created for visual words w^v.

We use the method in [10] to estimate the Granger Causality between any visual point process $N_i^v(t)$ and any audio point process $N_j^a(t)$. Firstly, the spectral matrix of the above two point processes is defined as follows:

$$S(f) = \begin{bmatrix} S_{1,1}(f) & \cdots & S_{1,N_v}(f) \\ \vdots & \ddots & \vdots \\ S_{N_a,1}(f) & \cdots & S_{N_a,N_v}(f) \end{bmatrix}, \tag{2}$$

where elements represent the cross-spectrum between visual point process $N_i^v(t)$ and audio point process $N_j^a(t)$. We used the multitaper method [28] to estimate the spectral matrix. In that method, M data tapers $\{h_m\}_{m=1}^M$ are applied sequentially to the point processes $N_i^a(t)$ and $N_j^v(t)$, and the Fourier transform of $N_i^a(t)$ is taken as follows:

$$\begin{aligned} \widetilde{P}_i^a(f, m) &= \int_0^T h_m(t) \exp(-i2\pi f t)\, dN_i^a(t) \\ &= \sum_j h_m(t_j) \exp(-i2\pi f t_j). \end{aligned} \tag{3}$$

The Fourier transform of $N_j^v(t)$, which denotes $\widetilde{P}_j^v(f, m)$, can be computed as same as (3). Then, the spectral matrix element $S_{i,j}(f)$ is estimated as follows:

$$\widehat{S}_{i,j}(f) = \frac{1}{2\pi MT} \sum_{m=1}^{M} \widetilde{P}_i^a(f, m)\, \widetilde{P}_j^v(f, m)^*. \tag{4}$$

For the time series of multivariate point processes $N^v(t)$ and $N^a(t)$, we adopt the autoregressive model to fit the data. The above $\widehat{S}_{i,j}(f)$ is then factorized as follows:

$$S_{i,j}(f) = H_{i,j}(f)\, \Sigma_{i,j} H_{ij}(f)^*, \tag{5}$$

where $H_{i,j}(f)$ is the transfer function determined by the coefficient matrix of the autoregressive model and Σ is the joint covariance of the error terms in the autoregressive model. Finally, the Granger Causality from $N_i^a(t)$ to $N_j^v(t)$ is then estimated by the method developed in [29] to

$$\begin{aligned} &G_{N_i^a \to N_j^v}(f) \\ &\quad = \ln\left(\frac{S_{jj}(f)}{S_{jj}(f) - \left(\Sigma_{ii} - \Sigma_{ji}^2/\Sigma_{jj}\right) \left|H_{ji}(f)\right|^2} \right), \end{aligned} \tag{6}$$

where f is all frequencies.

Notice that Granger Causality from $N_i^a(t)$ to $N_j^v(t)$ is not always equal to Grange Causality from $N_j^v(t)$ to $N_i^a(t)$ due to the directionality. Similarly, the Granger Causality from $N_j^v(t)$ to $N_i^a(t)$ is defined as follows:

$$\begin{aligned} &G_{N_i^a \leftarrow N_j^v}(f) \\ &\quad = \ln\left(\frac{S_{ii}(f)}{S_{ii}(f) - \left(\Sigma_{jj} - \Sigma_{ij}^2/\Sigma_{ii}\right) \left|H_{ij}(f)\right|^2} \right). \end{aligned} \tag{7}$$

Then the value of the Granger Causality between audio words and visual words is defined as the max value of two directions:

$$C_{a \leftrightarrow v}(i, j) = \max\left(G_{N_i^a \to N_j^v}(f), G_{N_i^a \leftarrow N_j^v}(f) \right). \tag{8}$$

4.2. Construction Audio-Visual Graph with Temporal Attribution. For all of the training videos, we extracted the visual and audio features in Section 3 and then form visual words w^v and audio words w^a by k-mean cluster method. In this section, we then define a weighted undirected graph $G = (V, E, w)$ to describe the causality between each word. Here V is the set of vertices which are represented as follows:

$$V = \{w^a, w^v\} = \left\{w_1^a, w_2^a, \ldots, w_{N_a}^a, w_1^v, w_2^v, \ldots, w_{N_v}^v\right\}. \tag{9}$$

Each node in V corresponds to a visual word or an audio word.

The set E is defined to measure the concurrence relationship between each word in V. The concurrence relationship between each word can be classified into three types, such as the relationship between visual words, the relationship between audio words, and the relationship between audio words and visual words. The Granger Causality between audio words and visual words is defined as (8). Similarly, the Granger Causality of the other types can be written as follows:

$$\begin{aligned} C_{v \leftrightarrow v}(i, j) &= \sum_f G_{N_i^v \to N_j^v}(f), \quad \forall i \neq j, \\ C_{a \leftrightarrow a}(i, j) &= \sum_f G_{N_i^a \to N_j^a}(f), \quad \forall i \neq j. \end{aligned} \tag{10}$$

In order to reduce the computation cost, we used a statistic threshold to discover the causal relationship in the Granger Causality matrix. Here we adopted three different thresholds $\mathrm{Th}_{a\leftrightarrow a}$, $\mathrm{Th}_{a\leftrightarrow v}$, and $\mathrm{Th}_{v\leftrightarrow v}$ for matrix $C_{a\leftrightarrow a}, C_{a\leftrightarrow v}, C_{v\leftrightarrow v}$, respectively. The value of Granger Causality scores that is less than the given threshold is regarded as a nontemporal relationship. After that, the score values that are larger than then threshold are normalized.

Based on above analysis, the weight w_{ij} of any edge $e_{ij} \in E$ $(i \le j)$ of the undirected graph G is defined as follows:

$$w_{ij} = \begin{cases} C_{a\leftrightarrow a}(i,j), & i < j \le N_a \\ C_{a\leftrightarrow v}(j-N_a,i)^*, & N_a < i \le N_a + N_v,\ j \le N_a \\ C_{v\leftrightarrow v}(i-N_a, j-N_a), & N_a < i < j \le N_a + N_v, \end{cases} \tag{11}$$

where $C_{a\leftrightarrow v}(j-N_a,i)^*$ denotes the transpose of matrix $C_{a\leftrightarrow v}$.

4.3. High-Level Codewords Representation for Event Detection. For the audio-visual graph $G = (V, E, w)$ we constructed in Section 4.2, a greedy agglomerative graph-partitioning method [30] is adopted to extract low-level word groups. Given the partition of the vertex set V into K groups $V_k = \{V_1, \ldots, V_k\}$, the maximum intragroup similarity is defined as follows:

$$\mathrm{Assoc}(V_k) = \sum_{i=1}^{k} \frac{S(V_i, V_i)}{d(V_i)}, \tag{12}$$

where $S(V_i, V_i)$ denotes sum of the weight of all edges in subset V_i and $d(V_i)$ denotes the sum of degree of all the vertex in subset V_i.

We start hierarchical clustering based on an improved association matrix which is defined on each edge of the weighted graph. The element in the improved matrix is defined as follows:

$$\mathrm{Delta}(A,B) = \frac{2S(A,B)}{(d(A) + d(B))}, \tag{13}$$

where A and B denote the different cluster in the graph. Initially, A or B is any vertex in the graph G. In each stage of clustering, we select the vertex pair (A^*, B^*), which has the maximum element in the matrix Delta, to form the a larger cluster AB^*. Then, matrix Delta is updated by removing the row and column related to A^* and B^*; at the same time, new row and column which denote the cluster AB^* are inserted into matrix Delta. In order to continue the next iteration steps, the weight matrix S and improved matrix are updated as follows:

$$\begin{aligned} S(AB^*, v) &= S(A^*, v) + S(B^*, v), \\ \mathrm{Delta}(AB^*, v) &= \frac{S(AB^*, AB^*) + S(v,v) + 2S(AB^*, v)}{d(AB^*) + d(v)} \\ &\quad - \frac{S(AB^*, AB^*)}{d(AB^*)} - \frac{S(v,v)}{d(v)}. \end{aligned} \tag{14}$$

The problem of determining the number of cluster is important in graph partition. In this paper, we adopted an effective method to determine order selection after initial hierarchical clustering [30]. In each step of the clustering, we define a new metric Curv to describe the similarity of the partition. The value of Curv is defined as follows:

$$\begin{aligned} \mathrm{Curv}(k) = &\left(\mathrm{Assoc}(V_k^*) - \mathrm{Assoc}(V_{k-1}^*)\right) \\ &- \left(\mathrm{Assoc}(V_{k+1}^*) - \mathrm{Assoc}(V_k^*)\right), \end{aligned} \tag{15}$$

where V_k^* denotes the partition which has the maximum normalized association over the partition of vertex set V into K clusters. Then the number of cluster is defined as follows:

$$K^* = \arg\max_k \mathrm{Curv}(k). \tag{16}$$

In practice, the value of K^* can be obtained by (16) or be provided by the user.

Each cluster in the graph partition forms a low-level word group which contains the temporal causality patterns between audio and visual features in the videos. And all the low-level groups form a high-level audio-visual dictionary, which is represented as $HD = \{HD_1, \ldots, HD_k\}$. Each audio-visual HD_i is represented as the combination of the audio words subset hd_i^a and the visual words hd_i^v in those high-level codewords.

For a given video V_i, the extracted visual feature and audio feature should be mapped into new audio-visual groups and then generate a high-level dictionary-based feature representation. Here we adopted an average pooling principle to aggregate original feature. The bag of high-level words is defined as follows:

$$H_i^g(k) = \frac{\sum_{w_m^a \in hd_k^a, w_n^v \in hd_k^v} h_i^a(m) + h_i^v(n)}{N(hd_k^a) + N(hd_k^v)}, \tag{17}$$

where $N(hd_k^a)$ and $N(hd_k^v)$ represent the number of audio words and visual words in the high-level codeword hd_k, w_m^a denotes the mth audio words, w_n^v denotes the nth visual words, $h_i^a(m)$ denotes the value of the mth bin in the audio words histogram representation of video V_i, and $h_i^v(n)$ means the value of the nth bin in the visual words histogram representation of video V_i. As seen from (17), the bag of high-level words representation is for all training videos, which is represented as follows:

$$H^g = \{H_1^g, \ldots, H_i^g\}. \tag{18}$$

4.4. Video Event Detection Based on Multiple Kernel Learning. Multiple kernel learning frameworks have been intensively applied in video analysis [11, 16–18]. In this paper, we combine our high-level codewords into the common used simpleMKL algorithm [11]. Since our high-level codewords include the temporal causality between visual and audio words, it is very difficult to decide the optimal size of our codewords.

We adopt different size of codewords representation in simpleMKL framework. The simpleMKL framework is defined to solve the following optimization problem:

$$\begin{aligned} \min_{\{f_n,b,\xi,d\}} \quad & \frac{1}{2}\sum_n \frac{1}{d_n}\left\|f_n\right\|_{H_n}^2 + C\sum_i \xi_i \\ \text{st.} \quad & y_j \sum_n f_n\left(x_j\right) + y_j b \geq 1 - \xi_j \quad \forall j \\ & \xi_j \geq 0 \quad \forall j \\ & \sum_n d_n = 1 \quad d_n \geq 0, \ \forall n. \end{aligned} \tag{19}$$

Due to the diversity of consumer videos in practical application, only a few properly labeled training data is given. Recently, Xu et al. proposed an event detection method to solve the problem of unlabeled training data, which can discriminate the positive and negative exemplars by learning multirelevance level label [18]. The multirelevance levels learning problem is given as follows:

$$\min_{f^m, y^m \in Y^m} \sum_{m=1}^{M} \left(\left\|f^m\right\|^2 + C\sum_{i=1}^{n} \ell\left(f^m, x_i^m, y_i^m\right) \right). \tag{20}$$

The above learning problem can be reformulated as

$$\min_{d\in\mathcal{D}} \left\{ \max_{\alpha\in\mathcal{A}} -\frac{1}{2}\alpha^T \left(\sum_{n:\widehat{y}_n\in Y} d_n K \bigodot \widehat{y}_n \widehat{y}_n^T \right) \alpha \right\}. \tag{21}$$

We use the matrices $K \odot \widehat{y}_n \widehat{y}_n^T$ as the basic kernels in MKL problem.

5. Experiment and Discussion

5.1. Experimental Setup. In this work, we evaluated our proposed high-level codewords representation for event detection over the large scale Columbia Consumer Video Dataset [3], which contains 9,317 consumer videos from YouTube (210 hours in total). These consumer videos contain diverse content without postediting, meanwhile the original audio tracks of the consumer videos are preserved. All of videos are manually labeled to 20 semantic categories. As same as the setting in [3], we use the same 4,659 videos for training and the remaining 4,658 videos for testing.

All our experiments were performed on a server machine with Intel Xeon 2.4 GHz CPUs and 32 GB RAM by using a single thread. For performance evaluation, we use average precision (AP, the area under precision-recall curve) and mean average precision (MAP, mean average precision across all event categories) as our evaluation metric [3].

In order to demonstrate the effectiveness of our method, we systematically perform the following methods:

(1) Individual feature: we performed our experiments on the four features (SIFT, STIP, MFCC, and DTF); however we will only report the result of STIP and DFT.

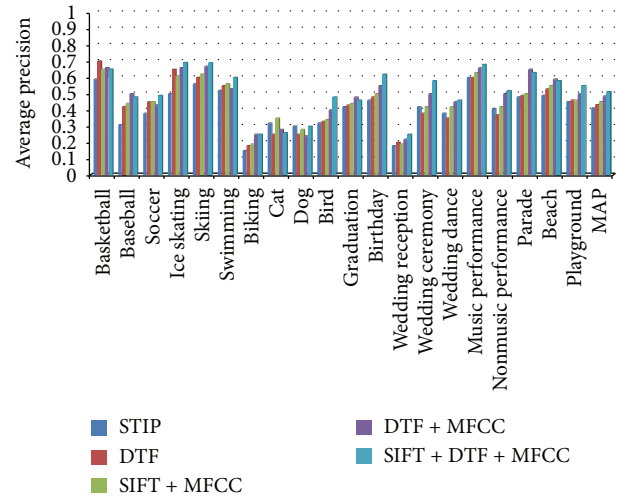

Figure 3: Performance of low-level features on CCV dataset.

(2) Early fusion: in order to evaluate the influence of audio information, we compared the performance of different manners of audio and visual combination, such as SIFT + MFCC, DTF + MFCC, and SIFT + DFT + MFCC.

(3) MKL based joint audio-visual codewords (MKL_AVC), where we use the joint audio-visual codewords in [21], especially, we just adopted the method of audio-visual graph construction in [21], and the method of graph partition is as described in this paper.

(4) MKL based high-level codewords (MKL_HLC): we used simpleMKL framework [11] to combine our high-level codewords based on temporal causality.

(5) Multilevel relevance labels and MKL based on high-level codewords (MLMKL_HLC): we used the multirelevance levels learning method in [18] to learn training label and then combined our high-level codewords to carry on event detection. In this experiment, each semantic category was labeled with R-level, and label R is for positive samples and label 1 is for negative samples. We fixed the parameter R as 4 for the multirelevance levels.

5.2. Performance of Low-Level Features. In the experiments of evaluating the performance of low-level feature, we trained a classifier for each semantic category by adopting one-versus-all x^2 kernel SVM, which has been proven by its outstanding performance for classifying BoW-based features. The AP and MAP results are shown in Figure 4.

As for the individual feature experiment, we can see that the four individual features have different advantages across different categories. In Figure 3, we present only the MAP of DTF and STIP, which achieved better performance in the four individual features. It can be observed that our results fall behind with the results in [3]. This because the bag-of-words histogram used here is normal, while the primary spatial layout representation is used in [3].

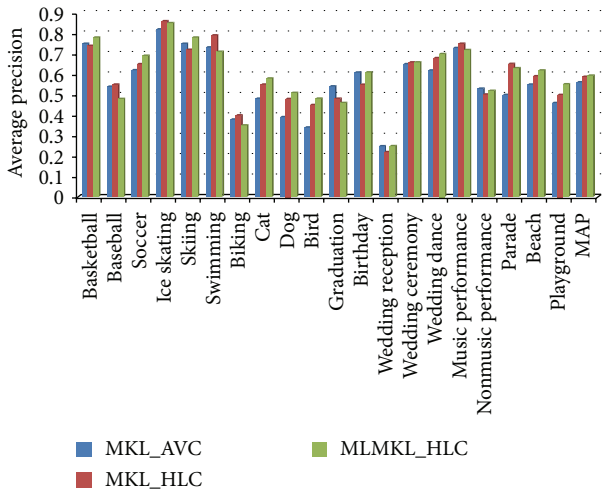

Figure 4: Performance of high-level features on CCV dataset.

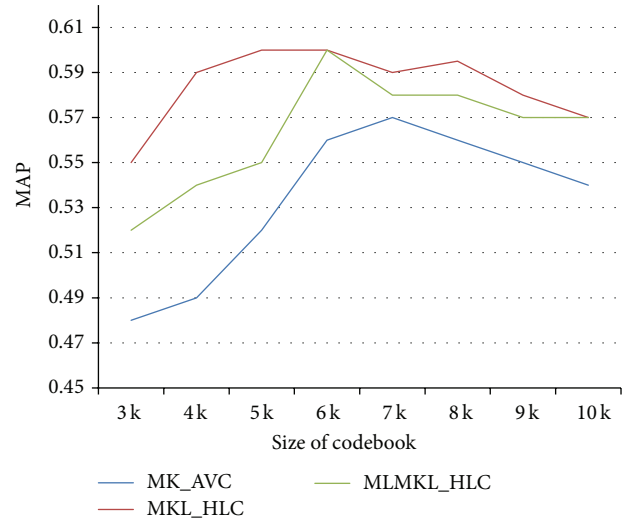

Figure 5: Performance of different codebook size for CCV dataset.

As for the early fusion of individual feature, we can see that the combination of audio and visual feature representation through early fusion improves the detection result. For example, the AP of all categories is obviously improved by the combination of the three single features (SIFT + DFT + MFCC), and the MAP is improved by nearly 10% on a relative basis.

5.3. Performance of High-Level Features. In the experiments of evaluating the performance of high-level feature, we compared our high-level codewords and the audio-visual codewords in [21]. Furthermore, we evaluated the performance of our high-level codewords under the simpleMKL framework in [11] and the multirelevance levels learning MKL framework in [18], respectively. According to the results of Section 5.2, we just incorporated SIFT, DTF, and MFCC into our high-level codewords.

As for the performance of methods based on high-level feature, we can see that the three methods (MKL_AVC, MKL_HLC, and MLMKL_HLC) outperform the methods based on individual feature and feature combination method. Such results were within our expectations because of the importance of the relationship between low-level codewords. Particularly, our proposed method (MKL_HLC) outperforms the baseline method MKL_AVC by nearly 3% in terms of MAP, which proves the effectiveness of our proposed method. For instance, on events "dog," our method (MKL_HLC) outperforms the individual feature STIP by 15% and outperforms the baseline method MKL_AVC by 9%. Besides, compared with the best baseline method MKL_AVC, our high-level codewords method achieves the highest relative performance gain on categories "birds" and "dogs." This may be because the emergence of visual object (bird or dog) often accompanies with the bark or warble. However, our method's performance is normal on the category "wedding reception," and this may be due to the large amount of background noise following people's actions. We also combine our high-level codewords into the multilevel relevance learning framework in [18]. We can see that MLMKL_HLC outperforms our MKL_HLC method by nearly 1% in terms of MAP, which indicate the effectiveness of the multirelevance levels learning in [18].

In general, we can expect a relative higher performance of the proposed method on other types of event which has obvious audio-visual association.

5.4. Codebook Analysis and Visualization. The different size of codebook can obviously impact the performance of event detection [21]. We hope each vocabulary can reflect a higher relativity between low-level words. Therefore, in the stage of high-level codewords representation (Section 4.3), the different number for order selection is manually selected. We compare the performance of different codebook sizes and different methods (MKL_AVC, MKL_HLC, and MLKL_HLC). The MAP performance is shown in Figure 5. We can see that the performance of the three methods gradually increases with the increasing codebook size. For this method, 6000 words seem to be the good choice for method MLMKL_HLC. The results of event detection in Section 5.3 are the performance using the best codebook size for our proposed method (MKL_HLC, MLMKL_HLC) and the baseline method (MKL_AVC).

We also compare the distribution of audio words and visual words in each high-level vocabulary of the two methods. For methods MKL_HLC and MKL_AVC, it is shown that the portion of audio-visual vocabulary, which contains both audio word and visual word, is found to be 45% and 34%, respectively. This proves that our high-level codewords can capture more association between audio word and visual word, compared to the bimodal words based on probability relationship in [21]. As indicated in the introduction of this paper, our high-level codewords are impactful for video events that contain audio-visual correlations. Figure 6 gives an example of this type of correlation. In the event "Birthday," the appearance of cake and candle often accompany with the birthday song, and then in the end of the song, there are some sounds of clapping and cheering. Figure 7 shows the high-level codeword of that video. Visual words in those high-level

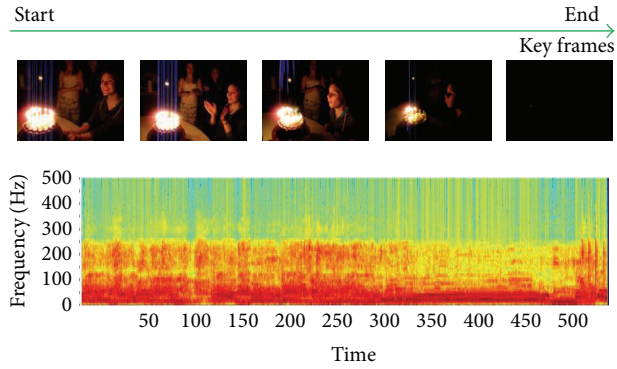

Figure 6: An example of audio-visual correlations in the event "Birthday" of CCV dataset.

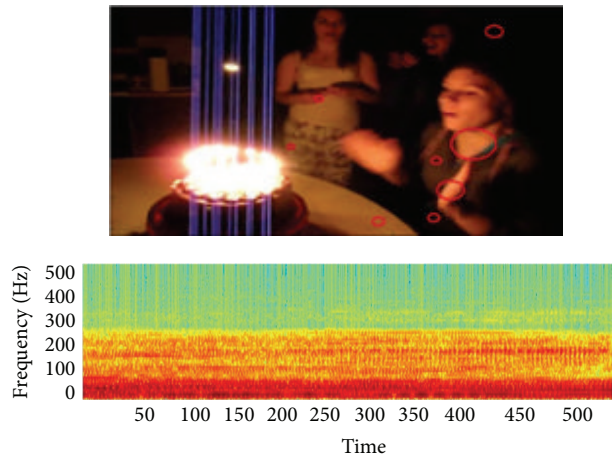

Figure 7: An example of high-level codeword in the event "Birthday" of CCV dataset.

Figure 8: An example of high-level codewords which include only visual words.

codewords are shown as sampled local points in the frame which are extremely close to the codebook vocabulary. Also the audio words in the high-level codewords are shown as the spectrogram of the sound over 500 s windows, where the MFCC features in that window are similar to the codebook vocabulary.

It is also observed that there are large numbers of vocabularies which contain only visual words or only audio words. The existence of these single channel vocabularies is reasonable because not every visual word is correlated to another audio word. Specifically, the audio words or the visual words in our method, which compose the single channel vocabularies, are also grouped together by the Granger Causality between them. We think that the effective single channel vocabularies which have the similar temporal patterns are also important cues for event detection. Figure 8 illustrates the effect of single channel vocabularies which only include visual words. For the video sequence of the

category "Wedding Dance," the visual words are shown in the first row, and different color circles are used to represent different visual words. The temporal high-level codewords in our method are shown in the second row. From Figure 8, we can see that the large majority of visual words produced by the hag action of the two characters are grouped into the same temporal group.

6. Conclusion

In this paper, we have introduced a high-level codewords representation framework for video event detection which can effectively utilize the low-level features in the video. By viewing the set of low-level words as the instantiation of multivariate point-process, we developed a Granger Causality graph to model the relationship between the low-level words of the videos. Then the graph is partitioned into low-level words groups which have the similar temporal patterns. Extensive experiments consistently show that the proposed high-level codewords representation outperforms the state-of-the-art multimodal fusion method. With these findings we can conclude that high-level codewords model representation will play important role in the future video event detection system. At the same time, advanced model representation will be worth to be intensively studied in the future to meet the practical application needs.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the Research Foundation of Education Bureau of Hunan Province, China (Grant no. 13C474).

References

[1] https://www.youtube.com/yt/press/statistics.html.

[2] T. Zhang, C. Xu, G. Zhu, S. Liu, and H. Lu, "A generic framework for event detection in various video domains," in *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)*, pp. 103–112, October 2010.

[3] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: a benchmark database and an evaluation of human and machine performance," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR '11)*, April 2011.

[4] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. G. Hauptmann, "Complex event detection via multi-source video attributes," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 2627–2633, June 2013.

[5] F. Wang, Z. Sun, Y. Jiang, and C. Ngo, "Video event detection using motion relativity and feature selection," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1303–1315, 2014.

[6] M. Tavassolipour, M. Karimian, and S. Kasaei, "Event detection and summarization in soccer videos using bayesian network and copula," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 2, pp. 291–304, 2014.

[7] A. Vahdat, K. Cannons, G. Mori, S. Oh, and I. Kim, "Compositional models for video event detection: a multiple kernel learning latent variable approach," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '13)*, pp. 1185–1192, December 2013.

[8] Y.-G. Jiang, S. Bhattacharya, S. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *International Journal of Multimedia Information Retrieval*, vol. 2, no. 2, pp. 73–101, 2013.

[9] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: an overview," in *Proceedings of the Issues in Visual and Audio-Visual Speech Processing*, 2004.

[10] A. G. Nedungadi, G. Rangarajan, N. Jain, and M. Ding, "Analyzing multiple spike trains with nonparametric Granger causality," *Journal of Computational Neuroscience*, vol. 27, no. 1, pp. 55–64, 2009.

[11] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[12] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA '05)*, pp. 399–402, ACM, November 2005.

[13] Y. Fu, L. Cao, G. Guo, and T. S. Huang, "Multiple feature fusion by subspace learning," in *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR '08)*, pp. 127–134, ACM, July 2008.

[14] S. Oh, S. McCloskey, I. Kim et al., "Multimedia event detection with multimodal feature fusion and temporal concept localization," *Machine Vision and Applications*, vol. 25, no. 1, pp. 49–69, 2014.

[15] P. Natarajan, S. Wu, S. N. P. Vitaladevuni, and X. Zhuang, "Multimodal feature fusion for robust event detection in web videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 1298–1305, IEEE, Providence, RI, USA, June 2012.

[16] Y. Yang, J. Song, Z. Huang, Z. Ma, N. Sebe, and A. G. Hauptman, "Multi-feature fusion via hierarchical regression for multimedia analysis," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 572–581, 2013.

[17] Z. Ma, Y. Yang, N. Sebe, and A. Hauptmann, "Multiple features but few labels? A symbiotic solution exemplified for video analysis," in *Proceedings of the International Conference on Multimedia (ACM MM '14)*, 2014.

[18] Z. Xu, I. W. Tsang, Y. Yang, Z. Ma, and A. G. Hauptmann, "Event detection using multi-level relevance labels and multiple features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 97–104, June 2014.

[19] N. Derbas and G. Quénot, "Joint audio-visual words for violent scenes detection in movies," in *Proceedings of the 4th ACM International Conference on Multimedia Retrieval (ICMR '14)*, pp. 483–486, April 2014.

[20] B. E. Ionescu, K. Seyerlehner, and I. Mironica, "An audio-visual approach to web video categorization," *Multimedia Tools and Applications*, vol. 70, no. 2, pp. 1007–1032, 2014.

[21] I.-H. Jhuo, G. Ye, S. Gao et al., "Discovering joint audio-visual codewords for video event detection," *Machine Vision and Applications*, vol. 25, no. 1, pp. 33–47, 2014.

[22] K. Prabhakar, S. Oh, P. Wang, G. D. Abowd, and J. M. Rehg, "Temporal causality for the analysis of visual events," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 1967–1974, IEEE, San Francisco, Calif, USA, June 2010.

[23] W. Jiang and A. C. Loui, "Audio-visual grouplet: temporal audio-visual interactions for general video concept classification," in *Proceedings of the 19th ACM International Conference on Multimedia (MM '11)*, pp. 123–132, December 2011.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[25] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, IEEE, Anchorage, Alaska, USA, June 2008.

[26] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 3169–3176, June 2011.

[27] L. Pols, *Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words*, Free University, Amsterdam, The Netherlands, 1996.

[28] A. T. Walden, "A unified view of multitaper multivariate spectral estimation," *Biometrika*, vol. 87, no. 4, pp. 767–788, 2000.

[29] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *Journal of the American Statistical Association*, vol. 77, no. 378, pp. 304–313, 1982.

[30] S. S. Tabatabaei, M. Coates, and M. Rabbat, "GANC: greedy agglomerative normalized cut for graph clustering," *Pattern Recognition*, vol. 45, no. 2, pp. 831–843, 2012.