Zero-shot Event Detection using Multi-modal Fusion of Weakly Supervised Concepts

Shuang Wu[†], Sravanthi Bondugula[‡], Florian Luisier[†], Xiaodan Zhuang[†] and Pradeep Natarajan[†*]

[†]Speech, Language and Multimedia
Raytheon BBN Technologies, Cambridge, MA
{swu,fluisier,xzhuang,pradeepn}@bbn.com

[‡]Department of Computer Science
University of Maryland, College Park, MD
sravb@cs.umd.edu

Abstract

Current state-of-the-art systems for visual content analysis require large training sets for each class of interest, and performance degrades rapidly with fewer examples. In this paper, we present a general framework for the zero-shot learning problem of performing high-level event detection with no training exemplars, using only textual descriptions. This task goes beyond the traditional zero-shot framework of adapting a given set of classes with training data to unseen classes. We leverage video and image collections with free-form text descriptions from widely available web sources to learn a large bank of concepts, in addition to using several off-the-shelf concept detectors, speech, and video text for representing videos. We utilize natural language processing technologies to generate event description features. The extracted features are then projected to a common high-dimensional space using text expansion, and similarity is computed in this space. We present extensive experimental results on the large TRECVID MED [26] corpus to demonstrate our approach. Our results show that the proposed concept detection methods significantly outperform current attribute classifiers such as Classemes [34], ObjectBank [21], and SUN attributes [28]. Further, we find that fusion, both within as well as between modalities, is crucial for optimal performance.

1. Introduction

Popular websites such as YouTube, Google images, and Flickr contain large volumes of image and video data from a multitude of consumer devices such as digital and cell-phone cameras. Technologies that can rapidly analyze such content and detect salient concepts and events have several compelling applications. Significant progress has been made in developing such technologies and the core of most state-of-the-art methods is based on the *bag-of-words* model [7]. Here, we first extract low-level features that capture salient gradient [22, 4], color [35], or motion [20, 37] patterns, project them to a pre-trained codebook in the same feature space, and then aggregate the projections to get the final image or video level feature vector. Classifiers, typically kernel support vector machines (SVM), are then trained using labeled data. This approach requires a large number of training examples for each class of interest and performance decreases rapidly as the training set size decreases.

In this paper, we study the problem of video classification using only a textual description of the events of interest, without exemplar videos pertaining to the events. This zero-shot framework, where we perform video classification with zero training samples, goes beyond traditional zero-shot problems such as described in [27], where an existing set of classes with training data is adapted to an unseen class. We pose this difficult problem of video classification as a retrieval task, where an event is described as a query defined by a set of concepts, e.g. the event "driving a car" described by the set of concepts "drive, car, road, person, face." We aim to retrieve videos that are most similar to the query, where the similarity score is treated as the confidence of the video belonging to that event.

Our approach to zero-shot learning is to first transform both video and query text to a high-dimensional concept space before computing similarity in that space. For the query, we apply text processing techniques to obtain a vector of salient words and phrases describing the event. For the video, we apply a bank of concept detectors to obtain a textual representation of the video using a vector of detection scores. Since we have no prior knowledge of the events

*This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

of interest, we need a very large set of generic concept detectors in order to provide semantic coverage of all possible queries. To address this challenge, we utilize multiple concept detectors from different modalities: visual features, including video concepts and multiple query fusion [1] of multiple features described in this paper, in addition to off-the-shelf detectors such as Classemes [34], ObjectBank [21] and SUN attributes [28]; audio information from concepts learned on low-level MFCC features; and text from video text and speech transcriptions.

Once we represent both query and videos as vectors of concept scores, we can compute similarities to retrieve relevant videos. A key challenge here is the mismatch between query and video concept vocabularies. We utilize a text expansion based method to project query and video concept vectors to a common high-dimensional concept space where they are compared, using the large text corpus Gigaword [14] to learn this projection matrix. Finally, we fuse retrievals from each of the features and modalities using a simple linear combination to exploit the complementary nature of the different modalities and concept vocabularies.

The paper is organized as follows: in Section 2, we discuss related approaches to similar problems. In Section 3, we present an overview of our zero-shot learning framework. Section 4 describes the features we extract from video and Section 5 outlines the combination of these features. We report experimental results in Section 6, and discuss our conclusions in Section 7.

2. Related Work

Extensive research has been performed in recent years on effective representation and classification of images and videos. The first step in most techniques is to extract *low-level features* from local spatial or spatio-temporal patches. Popular features include grayscale appearance features such as SIFT [22] and SURF [4], color features such as Color SIFT [35], and motion features such as STIP [20] and dense trajectories [37]. These typically extract thousands to millions of feature vectors per image or video. They are aggregated to a single fixed dimensional representation by a sequence of *coding* and *pooling* steps. Possible coding techniques include *Hard Quantization* [7], *Soft Quantization* [36], *Sparse Coding* [5] and *Fisher Vectors* [32], using a codebook trained in an unsupervised manner from a large set of feature vectors. The coded features are then aggregated, typically using average or max pooling, and classified typically using support vector machines (SVM).

While this approach has shown strong results given a large training set, performance degrades rapidly as the amount of training data decreases and the method does not generalize to previously unseen events. Only limited attention has been paid to this challenging problem and most existing approaches introduce an intermediate layer of semantic concepts, which are then used to describe novel classes. Semantic output codes (SOC) are proposed in [27] to extrapolate novel classes by utilizing a knowledge base of semantic properties of known classes. A large scale ontology is used in [31] to learn visual relationships between objects, while [30] uses knowledge transfer between object classes. An online incremental attribute based zero-shot learning approach is presented in [17], while a max-margin formulation is proposed in [15] for zero-shot multi-label classification where the label correlations on the training set differ significantly from the test set. A constrained optimization formulation that combines regression and knowledge transfer based functions has recently been proposed in [12].

All of these techniques rely on extrapolating from an existing set of classes and training data. The more difficult task of performing video retrieval and classification with no prior event knowledge or training data has been addressed only recently. In contrast to [8], we introduce several ways to generate a large visual and audio concept lexicon without prior knowledge of the event classes, and present a simple unified framework for effectively combining visual, audio, and textual information. While we are not able to benchmark our method against [8] since we do not have access to their concept lexicon or data partitions, our results in the TRECVID evaluation (Section 6.6) compare favorably to systems using similar approaches.

Video retrieval using semantic similarity has previously been explored in [2, 16]. However, these approaches focus on highly structured broadcast data, where a small 374 concept pool [2] can be adequate. In contrast, we focus on more challenging unconstrained web data where leveraging multiple modalities and larger concept banks is important to build a robust system. While [2, 16] both use a pre-defined concept ontology, we demonstrate the benefit of training in-domain detectors in a data driven manner by discovering concepts from free form text descriptions.

There has also been an increasing interest in joint modeling of text and visual features [3], which can then potentially be used to generate a text description of query images [13, 19, 38, 24] and videos [18, 9]. A large scale study of the relationship between semantic similarity of classes and confusion between them is presented in [11]. In [33], a large text corpus is used to learn a semantic space using word distributions and a separate model is trained for seen and unseen classes. However, given the training data limitations in our problem, we constrain our focus to attribute mappings produced using off-the-shelf features [34, 28, 21], novel concept banks developed with video-caption pairs similar to [24], and speech and video text output.

3. Zero-shot Learning Framework

Figure 1 displays an overview of our multi-modal zero-shot learning approach, which involves applying C different

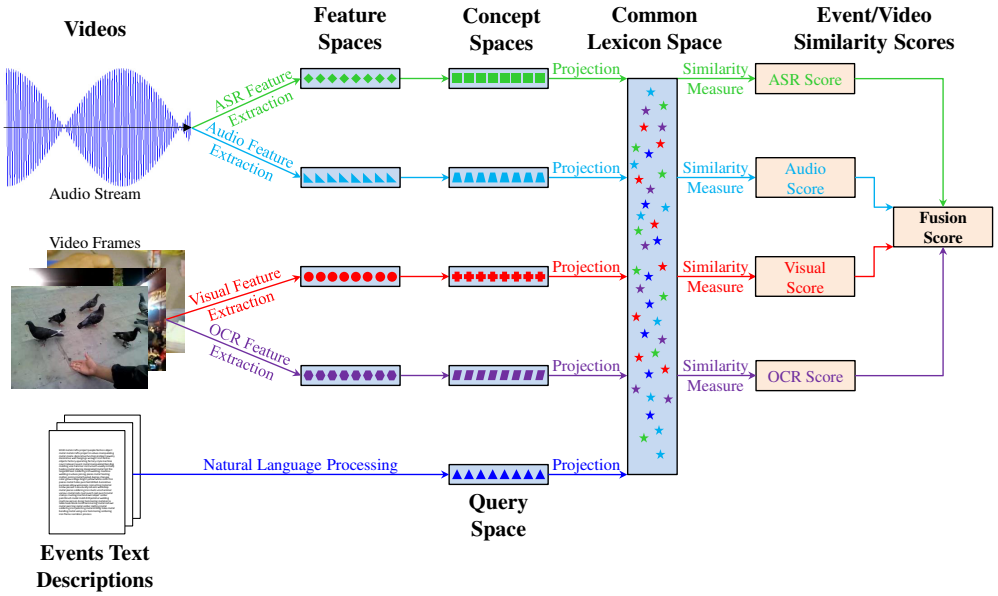

Figure 1. Overview of the proposed multi-modal zero-shot learning approach.

concept banks on each video v. Let $L = \{c_{l_1}, \ldots, c_{l_K}\}$ be a lexicon defined by K concepts $c_{l,k}$ for concept bank $l \in [1, \ldots, C]$. Each concept bank provides a K-dimensional vector of detection scores $\mathbf{d}_v = [d_{l_1} \ldots d_{l_K}]^T$ for each video $v = 1 \ldots V$, that is ℓ_2-normalized; i.e., $\|\mathbf{d}_v\|_2 = 1$. Given a query $Q = \{c_{q_1}, \ldots c_{q_N}\}$ defined by N concepts $c_{q,n}$, we aim to retrieve videos that are similar to the query.

3.1. Basic Similarity Computation

We first present a direct model to measure video-query similarity. In this model, we compute the similarity score $S_Q(v)$ between a query Q and a video v as a sum of the concept scores of the lexicon that match the query concepts:

$$S_Q(v) = \frac{1}{K} \sum_{k=1}^{K} d_{l_k} \mathbb{1}_Q(c_{l_k}) \qquad (1)$$

where $\mathbb{1}_Q(c_{l_k})$ is an indicator function of the presence of concept c_{l_k} in query Q.

This baseline system is very precise for efficient concept detectors. We expect the system to perform well when there is a large match between the query and video concepts, but lexicon coverage of the query will limit recall while noise in the video concept detections will degrade precision.

3.2. Expansion-based Similarity Computation

To address the issue of vocabulary mismatch between query and video, we use an alternative model to measure video-query similarity. In this model, concepts are expanded and projected to a common global concept space defined by the lexicon L. The goal is to propagate existing confidence scores to semantically similar concepts using the knowledge from a text corpus (like Gigaword) to estimate similarity. Let $G : (c_1, c_2) \longrightarrow s$ be a text model that measures the similarity $s \in [0, 1]$ between two concepts c_1 and c_2. Let an item I in the database be represented by a set of triplets describing the concept name, its confidence score, and its index in the lexicon L. The expansion-based projection method is given in Algorithm 1.

Algorithm 1 Expansion-based projection.

Given an item $I = \{(c_1, s_1, i_1), \ldots, (c_N, s_N, i_N)\}$.
Let $\mathbf{f} \in \mathbb{R}^K$ be the projected feature vector of item I for L.
Initialization: $f_k = 0$ for $k = 1 \ldots K$.
for each (c, s, i) in I **do**
  $f_i \leftarrow f_i + s$
  Find the top T similar concepts of c in G, given as
    $I_T = \{(c_1, s_1, i_1), \ldots, (c_T, s_T, i_T)\}$,
  where $s_t = G(c, c_t)$.
  Update the feature for the similar concepts:
  for each (c_t, s_t, i_t) in I_T **do**
    $f_{i_t} \leftarrow f_{i_t} + s \cdot s_t$
  end for
end for
Normalize the feature vector $\|\mathbf{f}\|_2 = 1$.

This algorithm obtains the projected vector $\mathbf{f}$ of an item I in two steps for each concept. The first step finds the top T similar concepts using the model G. The second step boosts the scores of the similar concepts for an item by the amount of similarity between the concepts. The final feature vector $\mathbf{f}$ is then normalized for comparison purposes.

Algorithm 1 is applied to expand both the query and

database concepts to a common lexicon space. Query concept confidences are given as 1, while database concept confidences are given by the output of the concept detectors. Once the expanded feature vectors $\mathbf{f}_Q \in \mathbb{R}^K$ representing the query Q and $\mathbf{f}_v \in \mathbb{R}^K$ representing the video v have been obtained, the similarity between the query Q and the video v is computed as

$$S_Q(v) = \mathbf{f}_Q^T \mathbf{f}_v. \quad (2)$$

Note that other similarity measures may also be considered (e.g., Laplacian or RBF kernels), although in our experiments we find that (2) has the best performance.

4. Video Feature Extraction

Since existing concept banks are generally trained on out of domain data and may not contain a large enough vocabulary to cover possible queries, we propose multiple methods to rapidly learn new concept detectors with easily collected data from readily available in-domain and web sources.

4.1. Weakly Supervised Concepts (WSC)

We train a set of WSCs for concept detection in videos using the following steps:

4.1.1 Data Collection and Concept Discovery

We collect a set of videos with free-form text descriptions of their content. Such data is widely available online in websites such as YouTube and also in the *research* set of the considered TRECVID MED dataset. We apply standard natural language processing (NLP) techniques to clean up the annotations, including removal of common stop words and stemming to normalize word inflections. The remaining vocabulary is taken as our concept dictionary.

4.1.2 Low-level feature extraction

For each video in the collected corpus, we extract the following set of low-level visual and audio features:

D-SIFT [5]: This is a dense version of SIFT where, instead of detecting interest points, the 128-dimensional feature vectors are extracted at uniformly-sampled locations covering the whole image. D-SIFT typically generates $3\times$ the number of points produced by SIFT [22] and has been shown to outperform SIFT for image classification [5].

Dense Trajectories (DT) [37]: This feature represents the video using dense optical flow trajectories. Histogram of oriented gradients (HoG) and motion boundary histograms (MBH) are extracted from the local spatio-temporal neighborhood of each track to capture salient appearance and motion patterns respectively.

MFCC [10]: These popular audio features are extracted from overlapping 29 ms frames at a rate of 100 frames per second. From each frame, we compute 14 mel-frequency warped cepstral coefficients. The resulting 45-dimensional feature vector captures the short-time spectral structure of the audio stream.

For each of the above low-level features, we first apply principal component analysis (PCA) to reduce the dimensionality and whiten the feature vectors. For each video, we then obtain a set $X = \{\mathbf{x}_t \in \mathbb{R}^D, t = 1 \dots T\}$ of T low-level low-dimensionality feature descriptors. We assume that these features are distributed according to a Gaussian mixture model (GMM) with diagonal covariance matrix:

$$p(\mathbf{x}_t|\Lambda) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2), \text{ for } t = 1 \dots T. \quad (3)$$

The GMM parameters

$$\Lambda = \{w_k \in [0,1], \boldsymbol{\mu}_k \in \mathbb{R}^D, \boldsymbol{\sigma}_k \in \mathbb{R}^D, k = 1 \dots K\}$$

are learned on a training set through maximum likelihood estimation. We then consider the Fisher vector encoding as proposed in [29] and represent each video by the normalized gradients of the GMM log-likelihood $\mathcal{G}_X^{\boldsymbol{\mu}_k} \in \mathbb{R}^D$ and $\mathcal{G}_X^{\boldsymbol{\sigma}_k} \in \mathbb{R}^D$ with respect to the Gaussian mean $\boldsymbol{\mu}_k$ and standard deviation parameters $\boldsymbol{\sigma}_k$, respectively. For $k = 1 \dots K$, these D-dimensional normalized gradients are defined as[1]

$$\mathcal{G}_X^{\boldsymbol{\mu}_k} = \frac{1}{T\sqrt{w_k}} \sum_{t=1}^{T} \gamma_k(\mathbf{x}_t|\Lambda) \left(\frac{\mathbf{x}_t - \boldsymbol{\mu}_k}{\boldsymbol{\sigma_k}} \right) \quad (4)$$

$$\mathcal{G}_X^{\boldsymbol{\sigma}_k} = \frac{1}{T\sqrt{2w_k}} \sum_{t=1}^{T} \gamma_k(\mathbf{x}_t|\Lambda) \left[\frac{(\mathbf{x}_t - \boldsymbol{\mu}_k)^2}{\boldsymbol{\sigma}_k^2} - 1 \right], \quad (5)$$

where the posterior probability

$$\gamma_k(\mathbf{x}_t|\Lambda) = \frac{w_k \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} w_l \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$$

is the soft assignment of the feature descriptor $\mathbf{x}_t$ to the k-th Gaussian cluster. The final Fisher vector is the concatenation of the K D-dimensional normalized gradients $\mathcal{G}_X^{\boldsymbol{\mu}_k}$ and $\mathcal{G}_X^{\boldsymbol{\sigma}_k}$, and is thus of dimension $2KD$.

4.1.3 Classifier Training

For each concept identified in Section 4.1.1, we collect all videos for which that concept occurs in the text caption, and utilize them as our positive training set, with the remaining videos considered as negatives. We then train RBF kernel-based support vector machine (SVM) classifiers using the Fisher vectors representing the videos. We train a set of concept detectors for each of the low-level features (D-SIFT, DT, MFCC) described in Section 4.1.2.

[1] Vector multiplications and divisions are element-wise operations here.

4.1.4 Weakly Supervised Concept Feature

Given a video, we produce a compact representation by concatenating the detection scores of our concept detectors. We use this feature vector for event detection and refer to this representation as *WSC*, for weakly-supervised concepts.

4.2. Concept Training using Web Data

In addition to the concept detectors trained using the research set described in Section 4.1.1, we also train detectors using data downloaded from the web. For each concept identified in Section 4.1.1, we downloaded the top 100 retrievals from Google images and thumbnails for the top 50 retrievals from YouTube. We then train WSCs with this data using the same approach as described in Section 4.1. We call the WSCs trained using the TRECVID research set, Google images and YouTube thumbnails as $\text{WSC}_{\text{TRECVID}}$, $\text{WSC}_{\text{Google}}$ and $\text{WSC}_{\text{YouTube}}$ respectively.

4.3. Concept Distance Features

We also introduce a novel concept distance (CD) based feature. Let C denote the set of concepts identified from the text annotations in Section 4.1.1. For each concept $c \in C$, let V_c denote the set of videos in the research set containing the concept. Let $\mathbf{x}_i$ denote the low-level feature based vector extracted for video i. Then, we compute the feature vector $\mathbf{y}_c$ for the concept c as:

$$\mathbf{y}_c = \frac{1}{|V_c|} \sum_{i \in V_c} \mathbf{x}_i. \tag{6}$$

Given a new video v and its low-level feature vector $\mathbf{x}_v$, we obtain the CD feature vector by computing the distance to each $\mathbf{y}_c$ in (6) and concatenating:

$$\mathbf{CD}_v = \left[\|\mathbf{x}_v - \mathbf{y}_1\|_2 \ \ldots \ \|\mathbf{x}_v - \mathbf{y}_{|C|}\|_2 \right]^T. \tag{7}$$

In our experiments, we use D-SIFT, DT and MFCC low-level feature vectors. The proposed feature vector builds on multiple-queries (MQ) [1] and the query expansion [6] based techniques proposed previously. While these approaches identify relevant videos at query time and use the retrievals to expand the concept set or training set, we use a static set of concept vectors $\mathbf{y}_c$ and compute distances at query time to these vectors.

4.4. Off-the-shelf Concept Detectors

We also test three off-the-shelf concept detectors that have been used in recent literature:

Classemes [34]: This is a bank of concept detectors trained on images. These were chosen using a large ontology of visual concepts. Given an image or a video frame, the application of all these detectors yields a 2,659-dimensional vector of detection scores.

ObjectBank [21]: Here, we use a spatial pyramid representation of images and produce detection confidence scores at different scales and spatial pyramids for each concept. The concept detectors are trained using linear SVMs and an image is represented by concatenating the detection scores of different concepts at different scales and spatial pyramids.

SUN Attributes [28]: The SUN attribute set contains detectors for 102 scene attributes that were specified using crowd sourced human studies.

We apply each of these concept detectors on a set of frames uniformly sampled from a video and then average the detection scores across the video to get the final video-level feature vector.

4.5. Automatic Speech Recognition (ASR)

We use GMM-based speech activity detection (SAD) and a hidden Markov model (HMM) based multi-pass large vocabulary ASR to obtain speech content in the video, and encode the hypotheses in the form of word lattices.

We first extract MFCC features from the audio stream. Then, the speech segments are identified by using a speech activity detection (SAD) system that employs two GMMs, for speech and non-speech observations respectively. The SAD model incorporates video clips with music content to enrich the non-speech model, in order to handle the heterogeneous audio in consumer video. Given the automatically detected speech segments, we then apply a large-vocabulary ASR system to the speech data to produce a transcript of the spoken content. The system is adapted from an ASR system trained on English Broadcast News, and updated with MED 2011 descriptor files [25], relative web text data, and the small set of annotated consumer video data. We evaluated the ASR model on a held-out set of 100 video clips and achieved a Word Error Rate (WER) of 35.8%. The system outputs not only the 1-best transcripts but also word lattices with acoustic and language model scores.

After basic processing to remove stop words and normalize word inflections, the word lattice posteriors are used to generate the concept score vectors used in the zero shot projection system.

4.6. Optical Character Recognition (OCR)

Our OCR system recognizes text in bounding boxes from a video text detector using an HMM-based multi-pass large vocabulary OCR system. Similar to our ASR system, word lattices are used to encode alternative hypotheses. We leverage a statistically trained video text detector based on SVM to estimate video text bounding boxes.

Text candidate regions are first selected using Maximally Stable Extremal Regions (MSER) and filtered using an SVM with rich shape descriptors such as Histogram of Oriented Gradients (HoG), Gabor filter, corners and geometrical features. Candidate regions are then grouped to

form word boundaries, and detected words are binarized and filtered before being passed to the HMM-based OCR system for recognition. The OCR system finds a sequence of characters that maximizes the posterior, by using glyph models (similar to the acoustic models in ASR), a dictionary and N-gram language models. The word precision and recall of our system measured on a small consumer video dataset is 14.7% and 37% respectively.

Since the video text content presents itself in various forms, such as subtitles, markup titles and in-scene text, it is much more challenging than conventional scanned document OCR. To address these challenges, we consider two versions of OCR: one which utilizes the dictionary and N-gram language model, and one which is character-based. While the language model corrects character-level transcription errors, it also introduce errors when falsely correcting out of vocabulary words. For the word model OCR output, we generate a concept score vector from the word lattice posteriors in the same way as ASR. For the character based model, we estimate word posteriors by smoothing character errors across adjacent video frames to produce a concept score vector. In our experiments we find the character model to be slightly better for video than the word model, as detailed in Section 6.

5. Fusion

State of the art systems for standard event detection with training data have shown fusion of multiple features and modalities to be crucial for improving performance [23]. Fusion is especially important for the zero-shot problem, due to the sparse occurrence of speech and video text content, as well as the limited vocabulary intersection between a given concept bank and query. While we do not have any training data on which to learn parameters for more sophisticated fusion methods, we find that simple score averaging works well to exploit the complementary information in various systems. We further see some benefit to manually increasing the weights of the higher precision ASR and OCR systems in fusion, and use a linearly weighted score combination for all fusion experiments below.

6. Experiments

We test our approach on the large collection of consumer web videos from the TRECVID MED 13 [26] dataset. The task is to retrieve videos containing one of 20 diverse high-level multimedia events, each described by a short text document of ∼250 words. The dataset provides a *research* set that contains ∼12,000 background videos and no exemplars of the events of interest. We use this research set to learn our $\text{WSC}_{\text{TRECVID}}$ and CD features. We report on the designated *MEDTest* set containing ∼25,000 videos. More details of the events and data partitions may be found in [26].

6.1. Comparison of Similarity Computation

Feature	Basic (MAP)	Expanded (MAP)
ASR	3.27%	**3.66**%
OCR (character)	4.43%	**4.72**%
CD^{MFCC}	**1.04**%	**1.04**%
$\text{WSC}^{\text{D-SIFT}}_{\text{YouTube}}$	3.42%	**3.48**%

Table 1. Mean average precision (MAP) comparison between basic (1) and expanded (2) query-video similarity computation for our single best ASR, OCR, audio, and visual features.

Table 1 compares the two methods of query-video similarity computation discussed in Section 3.1 and Section 3.2 for the best feature in each modality. We observe that expansion consistently improves over the simple approach. We observed similar gains from using projection based features in fusion, and thus we use the expansion-based approach in all experiments below.

6.2. Comparison of Visual Features

Feature	MAP	AUC
SUN [28]	0.48%	0.605
ObjectBank [21]	0.77%	0.592
Classemes [34]	0.84%	0.630
$\text{CD}^{\text{D-SIFT}}$	1.71%	0.770
CD^{DT}	2.28%	**0.779**
$\text{WSC}^{\text{D-SIFT}}_{\text{TRECVID}}$	1.92%	0.735
$\text{WSC}^{\text{DT}}_{\text{TRECVID}}$	2.76%	0.726
$\text{WSC}^{\text{D-SIFT}}_{\text{Google}}$	1.21%	0.543
$\text{WSC}^{\text{D-SIFT}}_{\text{YouTube}}$	**3.48**%	0.729

Table 2. Comparison of mean average precision (MAP) and area under the curve (AUC) for visual features.

In these experiments, we compare our proposed WSC and CD features to several off-the-shelf detectors. Table 2 summarizes our results. Here, $\text{WSC}^{\text{D-SIFT}}_{\text{YouTube}}$ refers to the weakly supervised concept features trained using D-SIFT features extracted on pre-downloaded YouTube thumbnails. Overall, the $\text{WSC}^{\text{D-SIFT}}_{\text{YouTube}}$ feature has the strongest performance, while the off-the-shelf detectors are significantly weaker than our proposed approaches. A possible reason for this is the large domain mismatch between the data used for training them and the video data. The same issue could explain the weaker performance of the $\text{WSC}_{\text{Google}}$ features compared to $\text{WSC}_{\text{TRECVID}}$ and $\text{WSC}_{\text{YouTube}}$ due to the domain mismatch between images and videos. Moreover, the CD features that are significantly faster to extract have comparable performance to the WSC features that require training expensive SVMs. Finally, the WSC and CD features detected using DT are stronger than the ones using D-SIFT.

6.3. Comparison of Audio Features

Feature	MAP	AUC
$WSC^{MFCC}_{TRECVID}$	0.76%	0.507
CD^{MFCC}	**1.04**%	**0.604**

Table 3. Comparison of mean average precision (MAP) and area under the curve (AUC) for audio features.

We compare the performance of our WSC and CD features trained using the audio MFCC features. Table 3 summarizes the MAP and AUC results. As observed, both of the audio features are weaker than the visual features.

6.4. Comparison of Language Features

Feature	MAP	AUC
ASR	3.66%	0.583
OCR (word)	4.30%	**0.636**
OCR (character)	**4.72**%	0.611

Table 4. Comparison of mean average precision (MAP) and area under the curve (AUC) for language features.

Table 4 compares the performance of our OCR and ASR systems. All the systems have higher MAP compared to the visual and audio features from Tables 2 and 3. However, note that the AUCs of many visual features outperform the language features. This is because although language content, when present, is a highly accurate source of information, its occurrence is sporadic, leading to low recall.

6.5. Comparison of Fusion Systems

Feature	MAP	AUC
ASR	3.66%	0.583
OCR	5.87%	0.642
Audio	1.04%	0.623
Visual (CD + WSC)	6.12%	**0.853**
Full	**12.65**%	0.733

Table 5. Comparison of mean average precision (MAP) and area under the curve (AUC) for fusion systems.

We fused each of the individual systems described above, both within each modality as well as across modalities. Table 5 compares the performance of the different fusion systems. Note that within the visual system, we found that off-the-shelf visual features did not improve the fused system, and only included our CD and WSC features. While none of the individual visual features is stronger than ASR or OCR, the visual system is the single strongest system after fusion, gaining $\sim$75% relative improvement over the single best visual system. The combined OCR system also outperforms the individual OCR systems, and the full system that combines all modalities more than doubles the performance of any individual modality as measured by MAP.

6.6. TRECVID Performance

The zero-shot event detection task was introduced as a pilot training condition as part of the TRECVID MED 13 evaluations. Independent evaluations were conducted by NIST on a blind $\sim$100000 video dataset, both for the same 20 events as in our previous experiments (*prespecified*), as well as for 10 new events given one week before the evaluation (*ad hoc*). Our zero-shot system achieved highly competitive scores for both prespecified and ad hoc conditions, placing among the top three out of 9 submissions. In particular, our consistent performance between prespecified and ad hoc events demonstrate the robustness of our event-independent approach to generalize to new queries.

7. Discussion and Conclusion

Only limited attention has been devoted to the task of video retrieval using only text queries. We present a systematic evaluation of our zero-shot framework for performing high-level multimedia event detection with no training data, given only text descriptions of the events of interest. Our findings and results on the large TRECVID MED dataset can serve as an initial baseline for this challenging task.

We present a general framework for zero-shot learning, that utilizes multiple multi-modal features to map a video to an intermediate semantic attribute space, which are then projected to a high-dimensional concept space using statistics learned on a large text corpus. Similarity between the attributes and a text query are computed in this space, and the scores computed from different attribute sets are combined to get the final score. We demonstrate the effectiveness of this approach for aligning disjoint vocabularies between query and various modalities.

We describe two simple but effective methods for rapidly training new concept detectors using in-domain as well as web data in the form of image/video with associated text descriptions. Detailed experimental results show that our concept detectors significantly outperform off-the-shelf detectors for zero-shot retrieval tasks. Exploiting the complementary nature of speech and video text as well as between different concept banks, we perform multiple rounds of fusion to produce a final system that is significantly better than any individual feature or modality.

References

[1] R. Arandjelović and A. Zisserman. Multiple queries for large scale specific object retrieval. In *BMVC*, 2012.

[2] Y. Aytar, M. Shah, and J. Luo. Utilizing semantic word similarity measures for video retrieval. In *CVPR*, pages 1–8. IEEE, 2008.

[3] K. Barnard, P. Duygulu, and D. A. Forsyth. Clustering art. In *CVPR*, pages 434–441, 2001.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *CVIU*, 110(3):346–359, 2008.

[5] Y. Boureau, F. Bach, Y. Le Cun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010.

[6] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 2007.

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

[8] J. Dalton, J. Allan, and M. Pranav. Zero-shot video retrieval using content and concepts. In *ACM Conference of Information and Knowledge Management*, 2013.

[9] P. Das, C. Xu, R. F. Doell, and J. J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, 2013.

[10] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *IEEE ASSP*, volume 28, pages 357–66, 1980.

[11] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, pages 71–84, 2010.

[12] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, 2013.

[13] A. Farhadi, S. M. M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. A. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, pages 15–29, 2010.

[14] D. Graff, J. Kong, K. Chen, and K. Maeda. English gigaword third edition. In *Linguistic Data Consortium, Philadelphia*, 2007.

[15] B. Hariharan, S. V. N. Vishwanathan, and M. Varma. Efficient max-margin multi-label classification with applications to zero-shot learning. *Machine Learning Journal*, 88(1):127–155, 2012.

[16] M. Y. Jung and S. H. Park. Semantic similarity based video retrieval. In *New Directions in Intelligent Interactive Multimedia Systems and Services-2*, pages 381–390. Springer, 2009.

[17] P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa. Online incremental attribute-based zero-shot learning. In *CVPR*, pages 3657–3664, 2012.

[18] N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama. Generating natural-language video descriptions using text-mined knowledge. pages 541–547, 2013.

[19] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR*, pages 1601–1608, 2011.

[20] I. Laptev. On space-time interest points. *IJCV*, 64(2-3):107–123, 2005.

[21] L.-J. Li, H. Su, E. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[23] P. Natarajan, S. Wu, S. N. P. Vitaladevuni, X. Zhuang, S. Tsakalidis, U. Park, R. Prasad, and P. Natarajan. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, 2012.

[24] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011.

[25] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, and G. Quenot. TrecVid 2011 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2011*. NIST, USA, 2011.

[26] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quenot. Trecvid 2013 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2013*. NIST, USA, 2013.

[27] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, December 2009.

[28] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758, 2012.

[29] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV*, volume 6314 of *Lecture Notes in Computer Science*, pages 143–156. Springer Berlin Heidelberg, 2010.

[30] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2011.

[31] O. Russakovsky and L. Fei-Fei. Attribute learning in large-scale datasets. In *ECCV*, Crete, Greece, September 2010.

[32] J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[33] R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. D. Manning, and A. Y. Ng. Zero-shot learning through cross-modal transfer. *CoRR*, abs/1301.3666, 2013.

[34] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *CVPR*, 2010.

[35] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. 32(9):1582–1596, 2010.

[36] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *IEEE PAMI*, 32(7):1271–1283, 2010.

[37] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*, 103(1):60–79, 2013.

[38] Y. Yang, C. L. Teo, H. D. III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP*, pages 444–454, 2011.