Semantic Concept Discovery for Large-Scale Zero-Shot Event Detection

Xiaojun Chang[1,2]**, Yi Yang**[1]**, Alexander G. Hauptmann**[2]**, Eric P. Xing**[3] **and Yao-Liang Yu**[3*]
[1]Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney.
[2]Language Technologies Institute, Carnegie Mellon University.
[3]Machine Learning Department, Carnegie Mellon University.
{cxj273, yee.i.yang}@gmail.com, {alex, epxing, yaoliang}@cs.cmu.edu

Abstract

We focus on detecting complex events in unconstrained Internet videos. While most existing works rely on the abundance of *labeled* training data, we consider a more difficult zero-shot setting where no training data is supplied. We first pre-train a number of concept classifiers using data from other sources. Then we evaluate the semantic correlation of each concept *w.r.t.* the event of interest. After further refinement to take prediction inaccuracy and discriminative power into account, we apply the discovered concept classifiers on all test videos and obtain multiple score vectors. These distinct score vectors are converted into pairwise comparison matrices and the nuclear norm rank aggregation framework is adopted to seek consensus. To address the challenging optimization formulation, we propose an efficient, highly scalable algorithm that is an order of magnitude faster than existing alternatives. Experiments on recent TRECVID datasets verify the superiority of the proposed approach.

1 Introduction

In multimedia event detection (MED), an event (class) of interest is specified and we must rank the (unlabeled) test videos so that positive examples (those containing the interested event) are ranked above the negative ones. As the first important step towards automatic categorization, recognition, search, indexing, and retrieval, video event detection has attracted more and more research attention in the computer vision and machine learning community [TRECVID, 2013]. With enough training data, a decent feature extraction module (*e.g.*, [Xu *et al.*, 2015; Laptev, 2005; Wang and Schmid, 2013; Sánchez *et al.*, 2013]) and a powerful statistical classification machine (*e.g.*, [Yan *et al.*, 2014; Wu *et al.*, 2009; Vahdat *et al.*, 2013]), one can achieve a reasonably good performance for event detection. However, MED faces the severe data-scarcity challenge: only very few, perhaps even none, positive training samples are available for some events, and the performance degrades dramatically once the number of positive training samples falls short. In some applications there is also the need to detect events that do not appear in the training phase at all.

Zero-shot learning [Lampert *et al.*, 2009; Larochelle *et al.*, 2008; Farhadi *et al.*, 2009; Palatucci *et al.*, 2009] is a recent remedy to the data-scarcity problem. Zero-shot event detection can be also interpreted as cross-media retrieval [Yang *et al.*, 2009] from text to videos. Without any training data, sample-based statistical approaches do not apply any more and we have to roll back to rule-based learning. The key observation is that many object classes can be relatively easily described as a composition of multiple middle-level attributes (concepts). For example, the *marriage proposal* event can be attributed to several concepts, such as ring (object), kissing (action) and kneeling down (action). Crucially, these concepts can be trained on other data sources and can be shared among many different events, including unseen ones. Based on this idea, events can be detected by inspecting the individual concept responses, even without any labeled training data [Dalton *et al.*, 2013; Habibian *et al.*, 2013; Wu *et al.*, 2014; Mensink *et al.*, 2014]. However, not all concepts are equally informative for detecting a certain event. We must also account for the inaccuracy of the concept classifiers, particularly when they are trained across different domains. Moreover, concept responses may have different scales and a naive way to aggregate them may not be the best strategy, potentially leading to the loss of precision. To the best of our knowledge, these issues have not been addressed in a principled manner for zero-shot event detection, and we intend to fill in this gap in this work.

The main building blocks of the proposed approach is illustrated in Figure 1. We begin with some background on zero-shot learning in §3.1, and then in §3.2 we learn a skip-gram language model [Mikolov *et al.*, 2013] to assess the semantic correlation of the event description and the pre-trained vocabulary of concepts. This step is carried out without any visual training data at all. Concepts are then ranked according to this semantic correlation. To further account for concept inaccuracy and discriminative power, in §3.3 we refine the concepts using either validation data or side information such as human knowledge. A highly selective subset of concepts are created for each event of interest, and we apply each of them to the (unseen) test videos, yielding a collection of confidence score vectors. These score vectors induce distinct rankings of the test videos hence we need to aggregate them effectively and

[*]To whom all correspondence should be addressed.

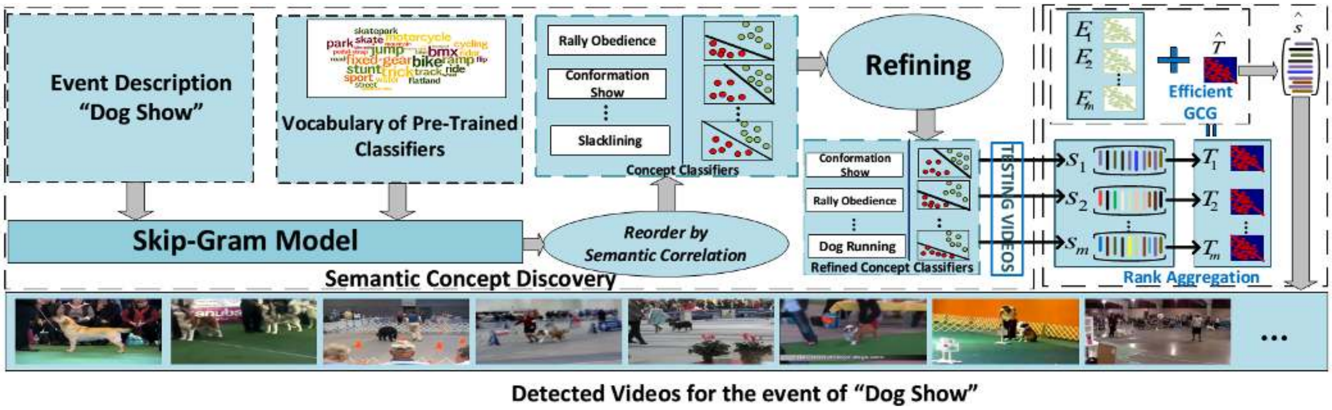

Figure 1: The proposed approach for large-scale zero-shot event detection (§3.1), illustrated on the particular "Dog Show" event. Discriminative concept classifiers are selected using the skip-gram language model (§3.2), followed by some further refinements (§3.3). The concept score vectors are combined through the rank aggregation framework (§3.4) and solved using the efficient GCG algorithm (§3.5).

efficiently, ideally also taking concept inaccuracy and score heterogeneity into account.

We leverage on the well-established rank aggregation framework [Dwork *et al.*, 2001; Gleich and Lim, 2011; Ye *et al.*, 2012]. Specifically, in §3.4 we first convert each score vector into a pairwise comparison matrix. Variations of the discriminative power of each concept in a certain feature space usually produce incomparable score vectors at different numeric scales, however, by removing the magnitude of the comparison matrix (*e.g*. taking its sign) we effectively enforce scale invariance. Besides, these comparison matrices are at best noisy perturbations of some groundtruth so we aim at finding their consensus under the robust Huber's loss [Huber, 1964]. Since the true pairwise comparison matrix is of low rank, the trace norm regularization is employed as a computational convenient proxy [Gleich and Lim, 2011; Chandrasekaran *et al.*, 2012]. Upon recovering the pairwise comparison matrix that is in consensus with the majority of those generated by individual concepts, we can easily recover a total ordering of all test videos, thus achieving zero-shot event detection. Lastly, since we have a large amount of test videos, in §3.5 we develop an efficient generalized conditional gradient algorithm [Zhang *et al.*, 2012, GCG] to solve our rank aggregation problem. The specialized algorithm is an order of magnitude faster than existing alternatives and works very well in our experiments (§4).

We summarize our main contributions as follows: 1). We apply data-driven word embeddings to correlate the middle-level concept classifiers with high-level event descriptions; 2). We fuse the concept score vectors through the framework of rank aggregation; 3). We develop the GCG algorithm to handle very large-scale test sets; 4). We conduct experiments on large-scale real-world video dataset, confirming the effectiveness of the proposed method for zero-shot event detection.

2 Related Works

A lot of past work in MED has been devoted to feature engineering and aggregating, *e.g*. [Natarajan *et al.*, 2012; Wang and Schmid, 2013; Li *et al.*, 2013; Oneaţă *et al.*, 2013]. More recent works have begun to explore the possibility of using middle-level semantic concept representations for event detection, *e.g*. [Ma *et al.*, 2013; Merler *et al.*, 2012; Habibian *et al.*, 2013]. Partial success along this vein has also been achieved in the zero-shot setting, *e.g*. [Habibian *et al.*, 2014; Wu *et al.*, 2014], but the concept score vectors are aggregated in a very limited way. Similar to us, [Dalton *et al.*, 2013; Mazloom *et al.*, 2013] also considered selecting more relevant concepts, albeit using explicit *labeled* training data. [Mensink *et al.*, 2014] exploited label co-occurrence for zero-shot learning.

Rank aggregation has a long history in social choice theory and we mention the influential work of [Dwork *et al.*, 2001]. Our work is inspired by the pairwise comparison framework of [Gleich and Lim, 2011] and the follow-up application in computer vision [Ye *et al.*, 2012]. However, both works did not consider event detection nor zero-shot learning. Besides, their algorithms cannot scale to the large datasets we are interested in here, see §3 and §4 for comparison.

3 The Proposed Method

Figure 1 illustrates the main building blocks of our proposed approach for zero-shot event detection. In a nutshell, we are given a large amount of unseen test videos and also the event description (such as a name), but without any labeled training data whatsoever (§3.1). Nonetheless, we need to rank the test videos so that positives (those contain the event of interest) are above negatives. With this goal in mind, we first associate a query event with some semantic *concepts* (attributes) that are pre-trained independently using other data sources (§3.2). After further pruning and refining (§3.3), we fuse the individual concept scores (on the *unseen* test data) by resorting to the well-established rank aggregation framework (§3.4). Finally, to handle a large amount of test data, we develop a very efficient optimization algorithm (§3.5).

3.1 Zero-shot learning

We first briefly recall the zero-shot learning problem, to provide context for the rest of the paper. Unlike traditional supervised learning tasks, zero-shot learning [Farhadi *et al.*, 2009; Lampert *et al.*, 2009; Larochelle *et al.*, 2008; Palatucci *et al.*, 2009] refers to the challenging scenario where we want to discriminate new classes that have no training data at all. Zero-shot learning appears frequently in practice because of the enormous amount of real-world object classes that are still constantly changing: it would be too time-consuming and expensive to get human annotated labels for each of them. Nevertheless, the crucial observation is that each object class can be semantically described as the composition of a set of *concepts*, *i.e.* middle-level interpretable attributes. For instance, the event *birthday party* can be described as the composition of "blowing candle", "birthday cake" and "applauding". Since concepts are shared among many different classes (events) and each concept classifier can be trained independently on datasets from other sources, zero-shot event detection can be achieved by combining the *relevant* concept classification sores, even in the absence of event labeled training data. In their pioneer work [Lampert *et al.*, 2009] largely relied on human knowledge to decompose classes (events) into attributes (concepts). Instead, we seek below an automated way to measure the similarity of an event of interest to the individual concepts.

3.2 Semantic correlation

Events come with textual side information, *e.g.* an event name or a short description. With the availability of a pre-trained vocabulary of concept (attribute) classifiers (see §4.2 for details) we can measure the semantic correlation between the query event and each individual concept. Since concept classifiers can be trained without any event label information, the semantic correlation makes it possible to share information between the concept space and the event space.

More precisely, we learn a skip-gram model [Mikolov *et al.*, 2013] using the English Wikipedia dump[1]. The skip-gram model infers a D-dimensional vector space representation of words by fitting the joint probability of the co-occurrence of surrounding contexts on large unstructured text data, and places semantically similar words near each other in the embedding vector space. Thus it is able to capture a large number of precise syntactic and semantic word relationships. For short phases consisting of multiple words (e.g. event descriptions), we simply average its word-vector representations. After properly normalizing the respective word-vectors, we compute the cosine distance of the event description and all individual concepts, resulting in a correlation vector $\mathbf{w} \in [0,1]^m$, where w_k measures the *a priori* correlation of the k-th concept and the event of interest. We further adjust the weight vector $\mathbf{w}$ below, taking their accuracy and predictive power into account.

3.3 Concept pruning and refining

In the previous section we have attached a weight $w \in [0,1]$ to each concept, indicating its similarity with the event of interest. These weights are further discounted (by a factor of say two) for a set of reasons: 1). Some concept classifiers may not be very reliable (low accuracy on test videos); 2) Some concepts, although relevant, may not be very discriminative (low predictive power); 3). Concepts trained on completely different domains (*e.g.* images) may be less useful for video event detection. If validation data is available (such is the case for TRECVID datasets), we can evaluate the above concerns about the concepts by computing their average precision on the validation data. Concepts with low precision (below a certain threshold) are then dropped. Non-discriminative concepts (such as "people") that appear almost in every video are also deleted. An added flexibility here is that we can easily incorporate human knowledge (if easily accessible), for instance, the concept "bike" will certainly be very useful for detecting the event *attempting a bike trick*, hence its weight w should be increased (by a factor of say two).

3.4 Rank aggregation

After constructing an appropriate subset of concepts for each event of interest, we use a rank aggregation strategy to conduct the detection. More precisely, suppose for event e we have selected m_e concepts, each with a weight $w_k \in [0,1], k = 1, \ldots, m_e$, to indicate its importance. Then, for n given (unseen) test videos, each concept classifier, say the k-th, generates a confidence score vector $\mathbf{s}^{(k)} \in \mathbb{R}^n$, which in turn induces a total ordering among the test videos:

$$\text{video } i \text{ ranked above video } j \iff s_i^{(k)} > s_j^{(k)}. \quad (1)$$

Different concepts usually generate distinct score vectors that induce different orderings on the test videos. However, as long as the majority of concepts are correct in ranking any pair of test videos, by aggregating the ranks to reach consensus among concepts it is still possible to recover the true ordering, as illustrated below.

A very straightforward approach is to use the weighted score vector

$$\mathbf{s} = \sum_{k=1}^{m_e} w_k \mathbf{s}^{(k)} \quad (2)$$

and its induced ordering as in (1). The weight vector $\mathbf{w}$ is incorporated to take into account the concept relevance. This is essentially the Borda count that is widely studied in social choice theory. As we show in the experiment, this approach does work reasonably well, but it suffers from the following drawbacks: 1). Each concept score vector is only a crude approximation of the groundtruth hence likely will make a few errors on some test videos, which the naive averaging strategy may not be able to correct; 2). Score vectors may not be on the same scale hence cannot be simply averaged. For instance, $\mathbf{s}^{(k)}$ and $10 \cdot \mathbf{s}^{(k)}$ both induce the same ordering but their influence to the average in (2) are completely different.

To address the above issues, we adopt the pairwise comparison framework in [Gleich and Lim, 2011]. Specifically, we convert the score vector $\mathbf{s}^{(k)} \in \mathbb{R}^n$ into a pairwise relationship matrix $T^{(k)} \in \mathbb{R}^{n \times n}$ in the following manner:

$$T^{(k)} = \mathbf{s}^{(k)} \cdot \mathbf{1}^\top - \mathbf{1} \cdot (\mathbf{s}^{(k)})^\top, \quad (3)$$

where $\mathbf{1}$ is the vector of all 1's (with appropriate dimension). Clearly, video i is ranked above video j according to the k-th concept classifier iff $s_i^{(k)} > s_j^{(k)}$ iff $T_{ij}^{(k)} > 0$. By simply taking the sign of the matrix $T^{(k)}$ we can enforce scale-invariance [Ye *et al.*, 2012].

[1] http://dumps.wikimedia.org/enwiki/

Algorithm 1: The GCG algorithm for the rank aggregation problem (4)

```
Set $T_0 = \mathbf{0}$, $s_0 = 0$.
for $t = 0, 1, \ldots$ do
    compute the gradient $G = \sum_{k=1}^{m_e} \nabla_T \mathsf{H}_{w_k,\gamma}(T_t - T^{(k)})$ ;
    find $(\mathbf{u}_t, -\mathbf{v}_t)$ as the leading singular vector pair of $G - G^\top$ ;          // Equation (10)
    $(a_t, b_t) \leftarrow \arg\min_{a,b} \sum_{k=1}^{m_e} \mathsf{H}_{w_k/\gamma}(a \cdot T_t + b \cdot (\mathbf{u}_t\mathbf{v}_t^\top - \mathbf{v}_t\mathbf{u}_t^\top) - T^{(k)}) + \lambda(a \cdot s_t + b)$   s.t.   $a \geq 0, b \geq 0$ ;
    $T_{t+1} \leftarrow a_t \cdot T_t + b_t(\mathbf{u}_t\mathbf{v}_t^\top - \mathbf{v}_t\mathbf{u}_t^\top)$ ;          // combining the old and new atoms
    $s_{t+1} \leftarrow a_t \cdot s_t + \frac{b_t}{2}(\|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2)$ ;          // Upper bound on trace norm
```

Now, given a collection of these pairwise comparison matrices $T^{(1)}, \ldots, T^{(m_e)}$, each corresponding to a relevant concept, we want to find the groundtruth T such that $T \approx T^{(k)}, k = 1, \ldots, m_e$, *i.e.* a consensus among the concepts. The key observation is that the matrix T constructed in (3) is a low-rank real asymmetric matrix (in fact rank 2, the lowest rank for a nonzero asymmetric matrix). This motivates us to consider the following optimization problem:

$$\min_{T, E^{(k)}} \sum_{k=1}^{m_e} w_k \|T + E^{(k)} - T^{(k)}\|_{\mathsf{F}}^2 + \gamma \|E^{(k)}\|_1 + \lambda \|T\|_{\mathrm{tr}}$$

$$\equiv \quad \min_{T} \quad \sum_{k=1}^{m_e} \mathsf{H}_{w_k,\gamma}(T - T^{(k)}) + \lambda \|T\|_{\mathrm{tr}}, \qquad (4)$$

where $\|\cdot\|_{\mathsf{F}}, \|\cdot\|_1, \|\cdot\|_{\mathrm{tr}}$ are the Frobenius norm, ℓ_1 norm (sum of absolute values), and trace norm (sum of singular values), respectively. The error matrix $E^{(k)}$ is introduced to accommodate possible errors made by the k-th concept and we penalize its ℓ_1 norm because hopefully the errors are few hence sparse. Other penalties, *e.g*. [Pan *et al.*, 2013], can also be easily adapted. As mentioned above, the consensus matrix T is of low rank hence we penalize its trace norm [Gleich and Lim, 2011], see also §3.5. Finally, the weight vector $\mathbf{w}$ is also incorporated to reflect the relevance of each concept with the event of interest. Interestingly, we can analytically eliminate the error matrices $E^{(k)}$ from (4), resulting in essentially the Huber's loss function [Huber, 1964]:

$$\mathsf{H}_{w,\gamma}(t) = \begin{cases} wt^2, & \text{if } |t| \leq \gamma/(2w) \\ \gamma|t| - \gamma^2/(4w), & \text{otherwise} \end{cases}, \qquad (5)$$

which is well-known to be robust against gross outliers.

After solving the consensus matrix T in (4), we can recover the underlying ranking using a two-step procedure. First, we recover its score vector $\mathbf{s}$ by solving the least squares problem

$$\min_{\mathbf{s}} \|\mathbf{s}\mathbf{1}^\top - \mathbf{1}\mathbf{s}^\top - T\|_{\mathsf{F}}^2, \qquad (6)$$

whose closed-form solution (up to a constant) is

$$\mathbf{s} = \tfrac{1}{n}\tfrac{T - T^\top}{2}\mathbf{1} = \tfrac{1}{n}T\mathbf{1}. \qquad (7)$$

Then, we recover the rank of each test video via (1).

The optimization problem (4) is convex and can be solved using a variety of ways. For similar problems, [Gleich and Lim, 2011] used the proximal gradient while [Ye *et al.*, 2012] employed the alternating direction method of multipliers (ADMM). However, both methods require a *full* singular value decomposition (SVD) on the matrix T in each iteration, leading to an $O(n^3)$ per-step complexity. For our application this is unbearable: The TRECVID 2013 dataset has $n = 23,954$ on the full set and $n = 12,388$ on the kindred set. Note that the matrix T is dense, therefore we cannot hope to reduce the cubic complexity by exploiting sparsity. We propose a different algorithm in the next section to bring down the time complexity to $O(n^2)$.

3.5 Optimization using GCG

We first motivate the trace norm regularizer in (4) under the atomic norm framework of [Chandrasekaran *et al.*, 2012]. Then we present a faster $O(n^2)$ time algorithm, known as generalized conditional gradient [Zhang *et al.*, 2012, GCG], for solving our main optimization problem (4).

Atomic norm For real asymmetric matrices, we define the atomic set as:

$$\mathsf{A} := \{\mathbf{u}\mathbf{v}^\top - \mathbf{v}\mathbf{u}^\top : \|\mathbf{u}\|_2 \leq 1, \|\mathbf{v}\|_2 \leq 1\}. \qquad (8)$$

These are the real asymmetric matrices of the lowest possible rank (0 or 2). Then, the induced atomic norm is given by

$$\|T\|_{\mathsf{A}} := \inf\{\rho \geq 0, T \in \rho \cdot \mathrm{conv}\mathsf{A}\}, \qquad (9)$$

i.e., we seek to decompose the asymmetric matrix T as a convex combination of low-rank "atoms". The atomic norm $\|T\|_{\mathsf{A}}$, as argued in [Chandrasekaran *et al.*, 2012], serves as a computational convenient proxy of the rank function. In fact, its dual norm is given by:

$$\|G\|_{\mathsf{A}}^\circ = \sup_{T \in \mathsf{A}} \langle T, G \rangle = \max_{\|\mathbf{u}\|_2 \leq 1, \|\mathbf{v}\|_2 \leq 1} \mathbf{u}^\top (G - G^\top)\mathbf{v}, \qquad (10)$$

i.e., the spectral norm of the asymmetric matrix $G - G^\top$. Dualizing again we immediately know the atomic norm $\|T\|_{\mathsf{A}} = \frac{1}{2}\|T\|_{\mathrm{tr}}$, hence justifying the trace norm regularization in (4).

Generalized conditional gradient (GCG) The GCG algorithm in [Zhang *et al.*, 2012] is suited for atomic norm regularization problems. We specialize it to our setting here. In each iteration, we first compute the gradient of the Huber loss

$$G = \sum_{k=1}^{m_e} \nabla \mathsf{H}_{w_k,\gamma}(T - T^{(k)}), \qquad (11)$$

and find the "new atom" $(\mathbf{u}, -\mathbf{v})$ in (10). Then we augment the previous atoms with the new atom:

$$T \leftarrow a \cdot T + b \cdot (\mathbf{u}\mathbf{v}^\top - \mathbf{v}\mathbf{u}^\top), \qquad (12)$$

MEDTest									KindredTest							
Prim	Selec	Bi	OR	SCD	SCD$_r$	Our	Our$_r$	Event ID	Prim	Selec	Bi	OR	SCD	SCD$_r$	Our	Our$_r$
5.3	4.9	4.7	7.6	6.4	8.3	12.5	16.3	E006	6.5	4.8	6.4	9.6	7.8	10.3	11.2	13.4
1.0	1.1	0.8	1.8	1.4	1.6	2.1	3.5	E007	1.2	1.2	1.3	1.1	1.1	1.4	1.7	2.1
18.4	23.0	9.0	31.9	28.1	33.1	37.8	43.4	E008	15.8	13.7	11.6	22.5	18.6	23.6	24.2	26.3
3.6	3.4	3.1	5.5	4.6	6.3	7.8	9.6	E009	1.5	1.5	2.6	2.2	2.1	2.4	2.6	3.1
0.9	0.9	0.8	0.9	0.7	0.9	1.1	1.5	E010	17.1	17.1	10.9	17.1	15.8	17.6	18.4	19.3
7.4	7.7	7.4	7.9	6.5	7.3	8.1	9.6	E011	64.2	62.6	64.2	66.9	61.2	68.4	71.2	79.4
19.8	21.9	19.3	22.4	18.4	24.3	33.2	35.9	E012	5.9	7.4	6.0	6.3	6.1	6.9	7.7	8.5
0.6	0.5	0.9	2.1	1.6	2.8	3.5	4.5	E013	1.0	0.8	1.1	4.3	3.8	4.5	5.9	6.8
1.1	1.2	0.9	2.5	1.7	3.4	4.1	5.8	E014	10.8	12.3	5.9	40.1	38.7	41.6	46.2	51.8
1.3	1.4	1.4	1.5	0.4	0.5	0.9	1.2	E015	24.6	30.0	30.9	27.3	25.4	29.5	32.5	35.4
1.1	1.1	0.6	2.0	0.9	1.3	2.5	3.3	E021	1.3	1.3	5.9	2.8	1.2	3.6	6.4	9.8
0.5	0.5	0.5	0.6	0.3	0.8	1.1	1.5	E022	10.2	7.6	10.2	23.9	21.6	25.1	28.4	31.7
0.1	0.1	0.3	0.1	0.4	0.3	0.5	0.8	E023	0.6	0.7	0.3	0.7	0.5	0.8	1.1	1.3
0.8	0.8	0.5	2.5	2.1	2.6	3.3	4.1	E024	0.3	0.2	0.2	0.5	0.3	0.6	0.6	0.9
0.1	0.1	0.2	0.2	0.2	0.4	0.6	0.7	E025	0.9	0.8	0.3	1.5	1.3	1.6	1.8	2.4
0.6	0.6	0.6	2.3	1.8	2.6	3.8	4.5	E026	4.2	6.7	4.9	6.7	5.8	7.1	7.4	8.8
14.0	14.2	13.9	14.6	12.3	16.5	19.9	21.3	E027	30.8	30.7	26.6	34.9	21.4	24.5	28.6	31.2
0.5	1.0	0.6	1.5	1.3	1.8	2.2	3.4	E028	0.7	1.6	0.6	0.4	0.3	0.7	1.2	1.9
2.7	3.1	3.1	11.0	10.2	12.4	16.4	19.8	E029	4.2	5.7	9.1	15.4	14.2	16.3	17.2	18.7
0.4	0.4	0.5	0.6	0.3	0.6	0.8	1.2	E030	7.4	7.4	9.6	9.9	4.3	6.3	7.8	9.2
4.0	4.4	3.5	5.9	4.9	6.4	8.1	9.6	mean	10.5	10.7	10.4	14.7	12.6	14.6	15.9	18.1

Table 1: Experimental comparisons for zero-shot event detection on TRECVID 2013. Mean average precision (mAP) is used as evaluation metric. Results are presented in percentage. Larger mAP indicates better performance. The proposed approach (4) outperforms related alternatives on both splits (Left: MEDTest. Right: Kindred.). Prim: Primitive, Selec: Selection, Bi: Bi-Concepts, OR: OR-Composite, SCD: Semantic concept discovery with (2), SCD$_r$: Refined concepts with (2), Our: The proposed approach, semantic concept discovery with (4), Our$_r$: The proposed approach, refined concepts with (4).

where $a, b > 0$ are found by line search. To avoid evaluating the trace norm (cubic time complexity) when performing line search, we use a similar variational upper bound as in [Zhang *et al.*, 2012]. Algorithm 1 summarizes the entire procedure. Following the argument in [Zhang *et al.*, 2012] we can prove that Algorithm 1 converges to an ϵ-optimal solution in at most $O(1/\epsilon)$ steps. The time complexity is reduced to $O(n^2)$ since the most time-consuming step is line 4, *i.e.*, computing the leading singular vector pair. Unlike *full* SVD which computes all n singular vector pairs, the leading singular vector pair costs only $O(n^2)$, an order of magnitude cheaper.

4 Experiments

In this section we conduct experimental evaluations of the proposed zero-shot event detection algorithm.

4.1 Synthetic

We first verify the efficiency of our GCG implementation against the ADMM algorithm in [Ye *et al.*, 2012]. We randomly generate a ground-truth pairwise comparison matrix T and corrupt it with i.i.d. Gaussian noise to generate $m = 5$ noisy $T^{(k)}$. We vary the size of T from $n = 1,000$ to $n = 10,000$ and terminate the rank aggregation algorithm when the difference between the estimated $\hat{T}$ and T falls below a threshold. As can be seen from Figure 2, the running time of [Ye *et al.*, 2012] increased sharply when the input size increases, due to its cubic per-step complexity. In comparison, the running time of our GCG implementation increased only modestly with the input size.

4.2 Experiment setup on TRECVID 2013

Vocabulary of concept classifiers: We pre-trained 1534 concept classifiers, using TRECVID SIN dataset (346 classes), Google sports (478 classes) [Karpathy *et al.*, 2014], ucf101 dataset (101 classes) [Soomro *et al.*, 2012] and YFCC dataset (609 classes) [YFC, 2014]. None of these datasets contains event label information. We extract the improved dense trajectory features (including HOG, HOF and MBH) and encode them with the fisher vector representation. Following [Wang and Schmid, 2013], we first reduce the dimension of each descriptor by a factor of 2 and then use 256 components to generate the Fisher vectors. On top of these features we trained the cascade SVM for each concept classifier.

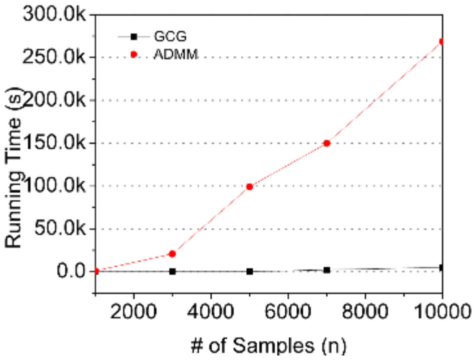

Figure 2: Efficiency comparison between GCG and ADMM.

Dataset description: We conduct experiments on the TRECVID Multimedia Event Detection (MED) 2013 corpus, including around 51,000 unconstrained videos on 20 pre-specified events. This is the largest publicly available unconstrained video corpus in the literature for event detection. The experiments are performed on two partitions of videos: MEDTest 2013 (around 27,000 videos) and KindredTest 2013 (around 14,000 videos). These two partitions

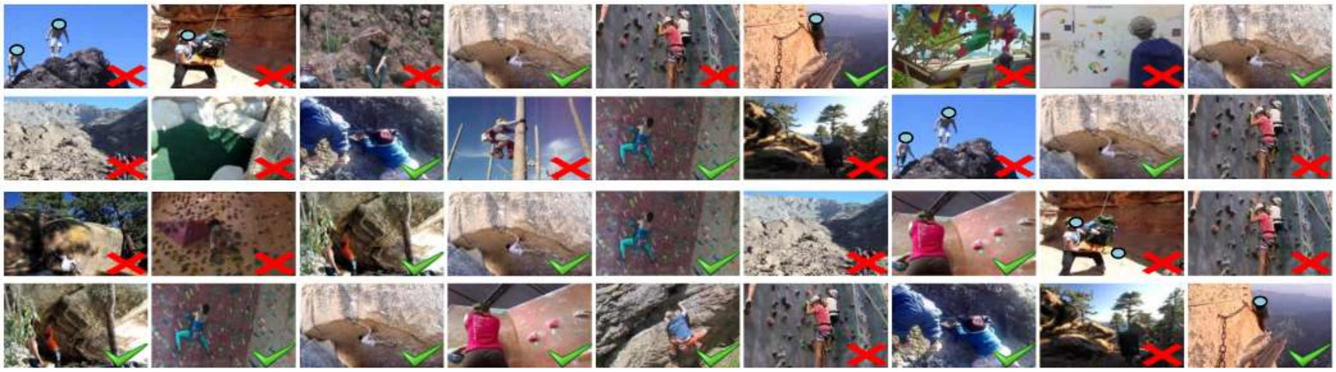

Figure 3: Top ranked videos for the event *Rock climbing*. From top to below are retrieved videos by selected concepts vocabulary, bi-concepts vocabulary, OR-composite concept discovery and the proposed method in Equation (4). True/false labels are marked in the lower-right of every frame.

are given with ground truth annotation (provided officially by NIST) at video level for 20 event categories, such as "Cleaning appliance", "Renovating home" and "Dog show". We use the official test split released by NIST, and following the standard in MED [TRECVID, 2013], we evaluate the classification accuracy using the mean Average Precision (mAP).

Compared Algorithms: We compare the proposed algorithm with the following baselines: 1). primitive concept vocabulary [Habibian *et al.*, 2014]. 2). bi-concepts vocabulary [Rastegari *et al.*, 2013]. 3). selected concepts vocabulary [Mazloom *et al.*, 2013], that is, a more informative subset of primitive concepts. 4). OR-composite concept vocabulary [Habibian *et al.*, 2014], *i.e.* combinations of concepts using Boolean logic operators. As we use the same data split, we directly quote the performances reported in the above references for fairness. We further compare against the simple Borda aggregation strategy in Equation (2), using either the discovered semantic concepts (SCD), or the refined semantic concepts in §3.3 (SCD$_\text{r}$).

4.3 Results on TRECVID 2013

The experimental results are shown in Table 1, from which we observe that the proposed method significantly outperforms primitive vocabulary with mAP of 0.163 vs 0.053 on MEDTest set and mAP of 0.134 vs 0.065 on KindredTest set. The proposed approach significantly improves on some events, such as *Birthday party (E006)*, *Flash mob gathering (E008)* and *Rock climbing (E027)*. For these events, the detection performance on MEDTest set is increased from 0.053 to 0.163, from 0.184 to 0.434 and from 0.140 to 0.213, respectively. By analyzing the discovered concepts of the proposed method, we find that their classifiers are very discriminative and reliable. For instance, for the event *Rock climbing* we discovered the concepts *Sport climbing*, *Person climbing vertically* and *Bouldering*, which are the most discriminative concepts for *Rock climbing* in the concept vocabulary. We also observe that the proposed method get slightly inferior performance on few events.

Our method is also comparable to bi-concepts vocabulary [Rastegari *et al.*, 2013], selected concepts vocabulary [Mazloom *et al.*, 2013] and OR-composite concept discovery [Habibian *et al.*, 2014]. As Table 1 shows, the proposed method performs better than the second best method, which is 0.096 vs 0.059 of mAP on MEDTest set. These experimental results indicate that the proposed method is capable of selecting the most reliable concepts and fusing the detection scores of different concept classifiers. Figure 3 illustrates the top retrieved results on the event *Rock climbing*. As we see, the videos retrieved by the proposed method are more accurate and visually coherent.

We also evaluate the performance of the simple Borda aggregation in Equation (2), on the discovered semantic concepts (SCD) and the refined concepts (SCD$_\text{r}$), respectively. The experimental results shown in Table 1 indicate that the performance of discovered semantic concepts is comparable to the baselines. SCD$_\text{r}$ generally performs better than SCD, which confirms that the refining step is capable of removing unreliable concepts and improving subsequent detection performance.

To further validate the effectiveness of the proposed approach, we remove the refining step but retain all other steps. From Table 1 it is clear that the performance becomes worse but is still competitive with the baselines. This indicates that unreliable concepts may deteriorate the detection accuracy, demonstrating further that better performance cannot always be guaranteed by more concepts.

Acknowledgment

This paper was partially supported by the 973 program (2012CB316400), partially supported by the ARC DECRA project, partially supported by the US Department of Defense, U. S. Army Research Office (W911NF-13-1-0277) and partially supported by the National Science Foundation under Grant No. IIS-1251187. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ARO, the National Science Foundation or the U.S. Government.

References

[Chandrasekaran *et al.*, 2012] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S.Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[Dalton *et al.*, 2013] Jeffrey Dalton, James Allan, and Pranav Mirajkar. Zero-shot video retrieval using content and concepts. In *CIKM*, 2013.

[Dwork *et al.*, 2001] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW*, pages 613–622, 2001.

[Farhadi *et al.*, 2009] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785, 2009.

[Gleich and Lim, 2011] David F. Gleich and Lek-heng Lim. Rank aggregation via nuclear norm minimization. In *KDD*, 2011.

[Habibian *et al.*, 2013] AmirHossein Habibian, Koen E. A. van de Sande, and Cees G. M. Snoek. Recommendations for video event recognition using concept vocabularies. In *ICMR*, 2013.

[Habibian *et al.*, 2014] AmirHossein Habibian, Thomas Mensink, and Cees G. M. Snoek. Composite concept discovery for zero-shot video event detection. In *ICMR*, 2014.

[Huber, 1964] Peter J. Huber. Robust estimation of a location parameter. *Annals of Statistics*, pages 73—101, 1964.

[Karpathy *et al.*, 2014] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[Lampert *et al.*, 2009] C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.

[Laptev, 2005] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[Larochelle *et al.*, 2008] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, 2008.

[Li *et al.*, 2013] Weixin Li, Qian Yu, Ajay Divakaran, and Nuno Vasconcelos. Dynamic pooling for complex event recognition. In *ICCV*, pages 2728–2735, 2013.

[Ma *et al.*, 2013] Zhigang Ma, Yi Yang, Zhongwen Xu, Shuicheng Yan, Nicu Sebe, and Alexander G. Hauptmann. Complex event detection via multi-source video attributes. In *CVPR*, 2013.

[Mazloom *et al.*, 2013] Masoud Mazloom, Efstratios Gavves, Koen E. A. van de Sande, and Cees Snoek. Searching informative concept banks for video event detection. In *ICMR*, 2013.

[Mensink *et al.*, 2014] T. Mensink, E. Gavves, and C.G.M. Snoek. COSTA: Co-occurrence statistics for zero-shot classification. In *CVPR*, pages 2441–2448, 2014.

[Merler *et al.*, 2012] Michele Merler, Bert Huang, Lexing Xie, Gang Hua, and Apostol Natsev. Semantic model vectors for complex video event recognition. *IEEE Transactions on Multimedia*, 14(1):88–101, 2012.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[Natarajan *et al.*, 2012] Pradeep Natarajan, Shuang Wu, Shiv Vitaladevuni, Xiaodan Zhuang, Stavros Tsakalidis, Unsang Park, Rohit Prasad, and Premkumar Natarajan. Multimodal feature fusion for robust event detection in web videos. In *CVPR*, 2012.

[Oneaţă *et al.*, 2013] Dan Oneaţă, Jakob Verbeek, and Cordelia Schmid. Action and event recognition with Fisher vectors on a compact feature set. In *ICCV*, pages 1817–1824, 2013.

[Palatucci *et al.*, 2009] Mark Palatucci, Dean Pomerleau, Geoffrey Hinton, and Tom M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

[Pan *et al.*, 2013] Yan Pan, Hanjiang Lai, Cong Liu, Yong Tang, and Shuicheng Yan. Rank aggregation via low-rank and structured-sparse decomposition. In *AAAI*, pages 760–766, 2013.

[Rastegari *et al.*, 2013] Mohammad Rastegari, Ali Diba, Devi Parikh, and Ali Farhadi. Multi-attribute queries: To merge or not to merge? In *CVPR*, 2013.

[Sánchez *et al.*, 2013] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob J. Verbeek. Image classification with the Fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

[Soomro *et al.*, 2012] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[TRECVID, 2013] TRECVID. Multimedia event detection. *http://www.nist.gov/itl/iad/mig/med13.cfm*, 2013.

[Vahdat *et al.*, 2013] Arash Vahdat, Kevin Cannons, Greg Mori, Sangmin Oh, and Ilseo Kim. Compositional models for video event detection: A multiple kernel learning latent variable approach. In *ICCV*, pages 1185–1192, 2013.

[Wang and Schmid, 2013] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[Wu *et al.*, 2009] Fei Wu, Yanan Liu, and Yueting Zhuang. Tensor-based transductive learning for multimodality video semantic concept detection. *IEEE Transactions on Multimedia*, 11(5):868–878, 2009.

[Wu *et al.*, 2014] Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Pradeep Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, 2014.

[Xu *et al.*, 2015] Zhongwen Xu, Yi Yang, and Alex Hauptmann. A discriminative cnn video representation for event detection. In *CVPR*, 2015.

[Yan *et al.*, 2014] Yan Yan, Elisa Ricci, Ramanathan Subramanian, Gaowen Liu, and Nicu Sebe. Multi-task linear discriminant analysis for multi-view action recognition. *IEEE Transactions on Image Processing*, 23(12):5599–5611, 2014.

[Yang *et al.*, 2009] Yi Yang, Dong Xu, Feiping Nie, Jiebo Luo, and Yueting Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *ACM Multimedia*, 2009.

[Ye *et al.*, 2012] Guangnan Ye, I-Hong Jhuo, Dong Liu, and Shih-Fu Chang. Robust late fusion with rank minimization. In *CVPR*, pages 3021–3028, 2012.

[YFC, 2014] The YFCC dataset. http://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67, 2014.

[Zhang *et al.*, 2012] Xinhua Zhang, Yaoliang Yu, and Dale Schuurmans. Accelerated training for matrix-norm regularization: A boosting approach. In *NIPS*, 2012.