They Are Not Equally Reliable: Semantic Event Search using Differentiated Concept Classifiers

Xiaojun Chang[1], Yao-Liang Yu[2], Yi Yang[1] and Eric P. Xing[2]
[1]Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney
[2]Machine Learning Department, Carnegie Mellon University
cxj273@gmail.com, yaoliang@cs.cmu.edu, yi.yang@uts.edu.au, epxing@cs.cmu.edu

Abstract

*Complex event detection on unconstrained Internet videos has seen much progress in recent years. However, state-of-the-art performance degrades dramatically when the number of positive training exemplars falls short. Since label acquisition is costly, laborious, and time-consuming, there is a real need to consider the much more challenging semantic event search problem, where no example video is given. In this paper, we present a state-of-the-art event search system without any example videos. Relying on the key observation that events (*e.g. *dog show) are usually compositions of multiple mid-level concepts (*e.g. *"dog," "theater," and "dog jumping"), we first train a skip-gram model to measure the relevance of each concept with the event of interest. The relevant concept classifiers then cast votes on the test videos but their reliability, due to lack of labeled training videos, has been largely unaddressed. We propose to combine the concept classifiers based on a principled estimate of their accuracy on the unlabeled test videos. A novel warping technique is proposed to improve the performance and an efficient highly-scalable algorithm is provided to quickly solve the resulting optimization. We conduct extensive experiments on the latest TRECVID MEDTest 2014, MEDTest 2013 and CCV datasets, and achieve state-of-the-art performances.*

1. Introduction

Multimedia event detection (MED) refers to the task of ranking a sequence of unseen videos according to their likelihood of containing a certain event, *e.g. birthday party*. Unlike concept/attribute (*e.g.* actions, scenes, objects) recognition, an event is a high level abstraction, possibly consisting of multiple concepts and spreading over the entire duration of long videos. For example, the *marriage proposal* event can be described by multiple objects (*e.g.* ring, faces), scene (*e.g.* in a restaurant), actions (*e.g.* talking, kneeling down) and acoustic concepts (*e.g.* music, cheering). Due to its apparent complexity and enormous utility in retrieval tasks, MED has drawn a lot of research attention in the computer vision and multimedia communities [14, 15, 29, 31, 9, 54, 12, 13].

A usual MED system first extracts low-level features from videos of interest to capture salient gradient [34, 5], color [51] or motion [52] patterns, and then encode these with a pre-trained codebook to get a succinct representation. With *labeled* training data, sophisticated statistical classifiers, such as support vector machines (SVM), are then applied on top to yield predictions. With enough labeled training examples, these systems have achieved remarkable performance in the past [29, 47, 31]. However, it is observed that performance decreases rapidly when the number of positive training exemplars falls short. Since in practice label acquisition is costly, laborious, and time-consuming, and also because of the constant need to handle new unseen events, the National Institute of Standards and Technology (NIST) initiated the zero-example search (0Ex for short) in TRECVID 2013 [1] and 2014 [2]. Promising progress [43, 53, 16, 21, 20, 11] has been made in this direction, but further improvement is still anticipated.

In this work we mainly focus on the semantic event search problem, where no example videos are provided for training whatsoever. Our system is built on the observation that an event is a composition of multiple mid-level concepts [30, 39, 10]. These concepts are shared among events and can be collected from other sources (not necessarily related to the event search task). We then train a skip-gram language model [37] to automatically identify the most relevant concepts to a particular event of interest. For example, the most relevant concepts for the *marriage proposal* event might be "face," "ring," "kissing," "kneeling down," *etc*. Such concept bundle view of event also aligns with the cognitive science literature, where humans are found to conceive objects as bundles of attributes [45]. The concept scores on the test videos are combined to yield a final ranking of the presence of the event of interest. However, this approach, as well as most existing works on semantic event

1063-6919/16 $31.00 © 2016 IEEE
DOI 10.1109/CVPR.2016.208

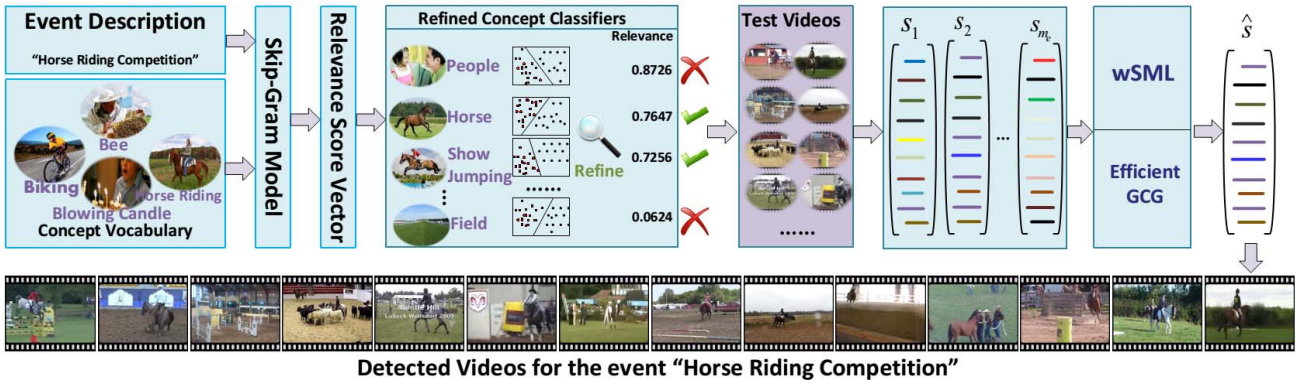


Figure 1: The proposed framework for large-scale semantic event search (§3), illustrated on the particular *horse riding competition* event. The relevance of concept classifiers to the event of interest are measured using the skip-gram language model (§3.2), followed by some further refinements (§3.3). To account for their reliability, the concept scores are combined through the warped spectral meta-learner (§3.6) and solved using the efficient GCG algorithm (§3.7).

search [43, 53, 16, 21, 20], ignore the fact that **not all concept classifiers are equally reliable**, especially when they are trained from other source domains. For example, "face" in video frames can now be reasonably accurately detected, but in contrast, the action "brush teeth" remains hard to recognize in short video clips. Consequently, a relevant concept can be of limited use or even misuse if its classifier is highly unreliable. Therefore, when combining concept scores, we propose to take their relevance, predictive power, and reliability all into account. This is achieved through a novel extension of the spectral meta-learner in [40], which provides a principled way to estimate classifier accuracies using purely *unlabeled* data. Figure 1 gives an overview of our entire system.

Contributions. To summarize, we make the following contributions in this work:

- To account for the unreliability of concept classifiers, we propose to use the warped spectral meta-learner to estimate the concept accuracies and combine them in a principled and purely unsupervised manner (§3.5 and §3.6).
- We provide an efficient implementation based on the recent generalized conditional gradient (§3.7), which is the key to conduct event search in large-scale video datasets.
- We conduct experiments on three real video datasets (MEDTest 2014, MEDTest 2013 and CCV_{sub}), and achieve state-of-the-art performances (§4).

2. Related works

Complex event detection on unconstrained Internet videos remains a very challenging task due to the large quality variations of Internet videos, the inherent complexity in event definitions, the limited number of positive training examples, and also the irregular appearance of the event in hour-long videos. Nevertheless, significant progress has been made in the past [29, 31, 47]. These approaches first extract low-level features (including appearance, motion, acoustic) from local spatial or spatial-temporal patches, and then aggregate them through coding [7, 41] and pooling [8, 49] to arrive at a succinct fixed-dimensional representation. Sophisticated supervised classifiers [33, 50] are then applied on top to yield predictions. With enough labeled training data, superb predictions can be achieved, but the performance of these *supervised* approaches drops dramatically when the number of positive examples decreases. Instead, we consider the more challenging semantic event search problem where *no* labeled exemplar data is provided.

Event detection with no training examples is called 0Ex for short. It mostly resembles a real-world video search scenario, where users search desired videos without providing any example video. Recent works have begun to explore intermediate semantic concepts [9], and achieved limited success on the 0Ex problem [43]. [53, 16, 21, 20] also considered selecting more informative concepts. However, none of these works considered the unreliability of the concept classifiers for event detection. [23] is closest to us in spirit, and considered unreliability for image classification without labeled training data. The limitation of their method is that they rely on the labeled validation data to account for attribute prediction unreliability. In our setting, no labeled validation data is provided. Hence, we cannot directly apply their algorithm to our problem.

We build on recent advances [17, 22, 40, 42] in estimating classifier accuracy using *unlabeled* data, which has received considerable attention in medical applications and more recently in crowdsourcing [44]. However, our work is the first to apply these techniques to the semantic event search problem, enhanced with a novel warping technique that significantly improves performance and an efficient implementation that allows scaling to real video datasets.

3. Semantic event search

In this work we mainly consider the semantic event search problem, where the learning algorithm is asked to rank unlabeled test videos according to their likelihood of containing a certain event of interest, for instance, *birthday party*. The significant challenge here is that we do not sup-

ply the learning algorithm with any labeled training data.

3.1. Concept classifiers

Without labeled training data, we can no longer train a supervised statistical classifier but resort to rule based learning. The key observation is that each object class can be described as the composition of a set of *semantic concepts*, *i.e*., middle-level interpretable attributes. For example, the event *marriage proposal* can be described as the composition of multiple objects (*e.g*., ring, faces), scene (*e.g*., in a restaurant), and actions (*e.g*., talking, kneeling down). Since concepts are shared among many different classes (events) and each concept classifier can be trained independently on datasets from other sources, semantic event search can be achieved by combining the *relevant* concept classification scores, even in the absence of event labeled training data. Different from the pioneer work [30], which largely relied on human knowledge to decompose classes (events) into attributes (concepts), we seek below an automated way.

3.2. Semantic concept relevance

Events come with short textual information, *e.g*., an event name or a short description. For example, the event *dog show* in the TRECVID MEDTest 2014 [2] is defined as "a competitive exhibition of dogs." We exploit this textual information by learning a relevance score between the event description and the pre-trained concept (attribute) classifiers. Since the concept classifiers are trained without any event label information, the relevance score makes it possible to share information between the concept space and the event space. More precisely, we pre-train a skip-gram model [37] 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 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 representation. After properly normalizing the respective word-vectors, we compute the cosine distance of the event description and all individual concepts, resulting in a relevance vector $\mathbf{w} \in [0,1]^m$, where w_k measures a priori relevance of the k-th concept and the event of interest. Similar approaches have appeared before in *e.g*. [36, 38, 53].

3.3. Concept pruning and refining

In the above we have introduced the relevance score vector $\mathbf{w} \in [0,1]^m$ that measures the similarity between the m concepts and the event of interest. We further prune and refine these weights for the following reasons: 1). Some concepts, although relevant to the event of interest, may not be very discriminative (low predictive power). For example, the concept *people* is relevant to the event *Birthday party*, but it appears almost in every video hence does not provide much discriminative power. 2). Some concepts may not be very reliable, possibly because they are trained on different domains. In the experiments, we use the (unlabeled) MED 2014 Research dataset[2] to crudely refine the concepts as follows: We first compute a similarity score between the concept names and the text description of each video in the research dataset, which acts as a *concept label*, *i.e*. the likelihood of each video to contain a particular concept. Then we run concept classifiers on each video in the research dataset, and use the aforementioned concept labels to compute the average precisions. Concepts with low precision or low predictive power (such as concept *people*) are then dropped. Importantly, our procedure does not require any manual annotation on the research dataset.

3.4. Combine the classifier ensemble

Suppose for event e we have selected m concepts[3], each with a weight $w_i \in [0,1],\ i = 1, \ldots, m$. Then, for any test video $\mathbf{v}$, the i-th concept classifier generates a confidence score $s_i(\mathbf{v}) \in [-1,1]$. Since different concept classifiers result in different confidence scores, we need a principled way to combine them, preferably also taking their relevance $\mathbf{w}$ into account. This can be treated as an ensemble learning problem, and there are many different ways to approach it. For instance, we can use each concept classifier i to induce a total ordering among n test videos, namely,

$$\text{video } k \text{ ranked above video } l \iff s_i(\mathbf{v}_k) > s_i(\mathbf{v}_l). \quad (1)$$

Then we can use rank aggregation techniques [18, 55] to combine the resulting ranks. A very intuitive and straightforward approach is to use the weighted score vector

$$\mathbf{s} = \textstyle\sum_{i=1}^{m} w_i \mathbf{s}_i \quad (2)$$

and its induced ranking as in (1). This is known as the Borda count in social choice theory, and has been explored in [4, 26, 36, 38] when no labeled training examples are given. In our later experiments, Borda works reasonably well. However, rank aggregation techniques can still be suboptimal, because the concept classifiers are obtained from other domains thus their accuracy on the test domain differs a lot. This motivates us to consider a recent approach that explicitly estimates the inaccuracy using *unlabeled* data.

3.5. Spectral meta-learning

Assuming for a moment that each score vector is binary, *i.e*. $s_i(\mathbf{v}) \in \{-1, 1\}$. We assume that the videos $\mathbf{v}$ are i.i.d.

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

[2] This adheres strictly to the NIST standard: "research set may be used for training concepts and assigning importance weights."

[3] Different events may use different concepts. For notational clarity, throughout we omit the dependence on the event e.

samples from an unknown distribution. The accuracy of the i-th concept classifier is defined as follows[4]:

$$p_i = \Pr(s_i(\mathbf{v}) = 1 | y = 1), \tag{3}$$

$$n_i = \Pr(s_i(\mathbf{v}) = -1 | y = -1), \tag{4}$$

$$\pi_i = (p_i + n_i)/2, \tag{5}$$

where y is the true event label of the test video $\mathbf{v}$, and $\pi_i \in [0, 1]$ is the average accuracy. Since we do not have labeled data, it is not immediately clear how we can estimate π_i.

The following assumption is standard for estimating classifier accuracy using *unlabeled* data [17, 22, 40]:

Assumption 1 (Conditional Independence)

$$\Pr(s_i(\mathbf{v}), s_j(\mathbf{v}) | y) = \Pr(s_i(\mathbf{v}) | y) \cdot \Pr(s_j(\mathbf{v}) | y) \tag{6}$$

In other words, given the label y, the classifiers make independent predictions. In our setting, the concept classifiers are trained from different sources, therefore the conditional independence assumption is reasonable.

Based on the conditional independence assumption, the following key observation is made in [40]:

Lemma 1 *Let $b = \Pr(y = 1) - \Pr(y = -1)$ be the class imbalance, $\mu_i = \mathbb{E}_{\mathbf{v}}(s_i(\mathbf{v}))$ be the mean prediction of the i-th concept classifier, and the population covariance matrix*

$$Q_{ij} = \mathbb{E}_{\mathbf{v}}[(s_i(\mathbf{v}) - \mu_i)(s_j(\mathbf{v}) - \mu_j)]. \tag{7}$$

Then, under the conditional independence assumption,

$$Q_{ij} = \begin{cases} 1 - \mu_i^2, & i = j \\ (2\pi_i - 1)(2\pi_j - 1)(1 - b^2), & i \neq j \end{cases}. \tag{8}$$

Crucially, from Lemma 1 we see that, except the diagonals, the population matrix Q arises from a rank-1 matrix, whose leading eigenvector $\mathbf{u}$ satisfies

$$u_i \propto (2\pi_i - 1). \tag{9}$$

This immediately leads to a principled way to estimate the accuracies π_i (up to a scale factor), since the covariance matrix Q can be easily estimated using unlabeled data. Consider the sample covariance matrix

$$\hat{Q}_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (s_i(\mathbf{v}_k) - \hat{\mu}_i)(s_j(\mathbf{v}_k) - \hat{\mu}_j), \tag{10}$$

where $\hat{\mu}_i = \frac{1}{n} \sum_{k=1}^{n} s_i(\mathbf{v}_k)$. Clearly, $\hat{Q}$ is an unbiased estimator of the population covariance matrix Q, and it can be shown that $\|\hat{Q} - Q\| = O_p(\frac{1}{\sqrt{n}})$. Therefore, for a large number of unlabeled data, we can estimate the accuracy π_i by solving the following problem:

$$\min_{R \succeq 0,\ \mathrm{rank}(R)=1} \quad \sum_{i \neq j} (\hat{Q}_{ij} - R_{ij})^2. \tag{11}$$

Note that it is important to exclude the diagonals of Q. Indeed, as shown in [40], the leading eigenvector of Q is a *biased* estimator of the accuracy π_i, and the bias depends on the number of classifiers m and the class imbalance b.

Unfortunately, (11) is a non-convex problem hence may be hard to solve. Instead, we turn to the following alternative, which uses the trace (since R is constrained to be positive semidefinite) as a convex surrogate for the nonconvex rank constraint:

$$\min_{R \succeq 0} \sum_{i \neq j} (\hat{Q}_{ij} - R_{ij})^2 + \lambda \operatorname{tr}(R). \tag{12}$$

The regularization constant λ controls the desired rank of the optimal solution. [40] proposed to solve (12) using generic semidefinite programming (SDP) toolboxes, which unfortunately do not scale very well. In Section 3.7 we will provide a much faster $O(m^2)$ time algorithm.

After solving R from (12), we extract the accuracy π_i from its leading eigenvector $\mathbf{u}$. Now the question is can we combine the classifiers more smartly by taking their accuracy into account? The answer is yes, and traces back to [17], which considered the maximum likelihood estimator:

$$y^* = \operatorname{sign}\left[\sum_{i=1}^{m} (s_i(\mathbf{v}) \log \alpha_i + \log \beta_i)\right], \tag{13}$$

$$\alpha_i = \frac{p_i n_i}{(1 - p_i)(1 - n_i)}, \quad \beta_i = \frac{p_i(1 - p_i)}{n_i(1 - n_i)}. \tag{14}$$

To get α and β from the accuracy π, [40] considered Taylor expansion of the MLE at the most inaccurate setting $p_i = n_i = 1/2$. This yields the spectral meta-learner (SML):

$$\hat{y} = \operatorname{sign}\left[\sum_{i=1}^{m} s_i(\mathbf{v})(2\pi_i - 1)\right] \approx \operatorname{sign}\left[\sum_{i=1}^{m} s_i(\mathbf{v}) u_i\right], \tag{15}$$

where recall that $\mathbf{u}$ is the leading eigenvector of the minimizer R of (12). Interestingly, the spectral meta-learner is essentially a weighted majority voting rule, where the weights come from the estimates of the accuracy. Intuitively, it gives more weight to classifiers whose estimated accuracy is high, and vice versa. We note that it is possible to construct the meta-learner using more sophisticated tensor approaches [22].

3.6. specialization and extension

In this section we specialize the spectral meta-learner above to our semantic event search framework.

Probabilistic classifiers. Recall that we obtain m concept classifiers from other domains and apply them to n unlabeled test videos, resulting in the score vectors $\mathbf{s}_i \in [-1, 1]^n, i = 1, \ldots, m$. The theory in section 3.5 requires $\mathbf{s}_i$ to be binary, but this can be easily addressed by treating each score vector $\mathbf{s}_i$ as *probabilistic* classifiers, namely, we classify the k-th test video as positive with probability $s_i(\mathbf{v}_k)$, independently of everything else. Under this interpretation we can still derive Lemma 1, the sample covariance $\hat{Q}$, and the spectral meta-learner as before, without the need of thresholding the score vectors.

[4] We implicitly assume that the scores are *positively* related to the label.

Warping functions. Next, we wish to incorporate the relevance vector $\mathbf{w}$ that we constructed in section 3.2 and refined in section 3.3. To see why this is desirable, let us first note that Lemma 1 applies to *any* classifiers, as long as they satisfy the conditional independence assumption. More precisely, for transformations f_i that do not depend on the unseen test video $\mathbf{v}$ or its unknown label y, we can consider the "warped" classifiers

$$t_i(\mathbf{v}) = f_i(s_i(\mathbf{v})), \quad i = 1, \ldots, m. \tag{16}$$

Clearly, the warped classifiers $\mathbf{t}$ are conditionally independent if and only if the original classifiers $\mathbf{s}$ are so. Therefore Lemma 1 still holds, and we can construct the sample covariance matrix

$$\hat{Q}^f_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(t_i(\mathbf{v}_k) - \hat{\mu}^f_i)(t_j(\mathbf{v}_k) - \hat{\mu}^f_j), \tag{17}$$

where as before $\hat{\mu}^f_i = \frac{1}{n}\sum_{k=1}^{n} t_i(\mathbf{v}_k)$. The spectral meta-learner for the warped classifiers is thus given as:

$$\hat{y}^f = \operatorname{sign}\Big[\sum_{i=1}^{m} f_i(s_i(\mathbf{v}))u_i\Big], \tag{18}$$

where $\mathbf{u}$ is the leading eigenvector of R, the minimizer of (12) where we use $\hat{Q}^f_{ij}$ instead.

Warped spectral meta-learner. Straightforward as it is, the extension using different warping functions f_i can lead to a significant performance improvement. This is because the accuracy of the spectral meta-learner $\hat{y}$ in (15) depends on the accuracies of the base classifiers s_i: SML is a smart way to combine the base classifiers, but we should not expect it to improve the accuracy much if the base classifiers are themselves near random. After all, garbage in garbage out. The warping functions f_i provide an extremely simple way to adjust the base classifiers. Since the relevance vector $\mathbf{w}$ we constructed in Section 3.2 provides a crude assessment of the relevance between the concept classifiers and the event of interest, we consider the following warped concept classifiers:

$$\mathbf{t} = (w_1 s_1, \ldots, w_m s_m), \tag{19}$$

although other warping functions can similarly be used. Intuitively, the weight w_i is the *a priori* co-occurrence frequency of the i-th concept and the event of interest while s_i is the confidence *likelihood* of detecting the i-th concept. As we will see in the experiments, this simple warping trick significantly improves the performance.

Few exemplars. The warped spectral meta-learner above can also be applied for few-exemplar event detection, where few (say 10) labeled training videos are provided. In this case, we can train an additional classifier (or few) using the provided labeled videos. Due to the small training size, the accuracy of the resulting supervised classifier is likely also low. We combine the supervised classifier with the concept classifiers but give it the maximum weight $w = 1$. Then we apply the warped spectral learner to get the final prediction.

Algorithm 1: The warped SML algorithm.

Construct concept classifiers $\mathbf{s}$ and relevance vector $\mathbf{w}$.
Apply warping $\mathbf{t} = (f_1(s_1), \ldots, f_m(s_m))$.
Assemble the sample covariance $\hat{Q}^f$.
Set $U_1 = \mathbf{0}$.
for $t = 1, 2, \ldots$ **do**
  $R \leftarrow U_t U_t^\top$;
  $G_{ij} \leftarrow \begin{cases} 0, & i = j \\ R_{ij} - \hat{Q}_{ij}, & i \neq j \end{cases}$;
  $\mathbf{u} \leftarrow$ leading eigenvector of $-G$;
  $(a_t, b_t) \leftarrow \arg\min_{a,b\geq 0} \sum_{i\neq j}(aR_{ij} + bu_iu_j - \hat{Q}_{ij})^2 + \lambda(a\operatorname{tr}(R) + b)$;
  $U_{\text{init}} \leftarrow (\sqrt{a_t}U_{t-1}, \sqrt{b_t}\mathbf{u})$;
  $U_t \leftarrow$ local minimizer of (23), initial with U_{init} ;
$\mathbf{u} \leftarrow$ leading eigenvector of R ;
Rank test videos using (18).

3.7. Optimization using GCG

Lastly, we provide a fast algorithm for solving the semidefinite program (12). This is crucial if we want to combine a large number of concept classifiers.

We use the generalized conditional gradient (GCG, *a.k.a* Frank-Wolfe) algorithm in [57, 10], with essential modifications to take the positive semidefinite constraint into account. In each iteration, GCG first computes the gradient

$$G = \nabla_R\Big[\textstyle\sum_{i\neq j}(R_{ij} - \hat{Q}_{ij})^2\Big]. \tag{20}$$

Then it finds a rank-1 update

$$\mathbf{u} = \operatorname*{argmin}_{\|\mathbf{z}\|_2\leq 1} \mathbf{z}^\top G\mathbf{z}, \tag{21}$$

which amounts to the leading eigenvector of $-G$. This step takes into account the trace regularizer, and is essentially its dual operator (spectral norm). Finally, GCG augments the previous iterate R with the new rank-1 update:

$$R \leftarrow a \cdot R + b \cdot \mathbf{u}\mathbf{u}^\top, \tag{22}$$

where the positive coefficients a, b are found by line search.

To accelerate convergence, we consider the following smooth *unconstrained* problem:

$$\min_U \sum_{i\neq j}((UU^\top)_{ij} - \hat{Q}_{ij})^2 + \lambda\|U\|^2_{\mathsf{F}}, \tag{23}$$

which, unlike the original problem (12), is nonconvex. But we can combine the global GCG algorithm with a local fast solver for the nonconvex problem (23). The intuition is that both (12) and (23) share the same set of global minimizers, and by combining them we gain both global optimality and local fast convergence, especially because the latter nonconvex problem has no constraint at all. We summarize the

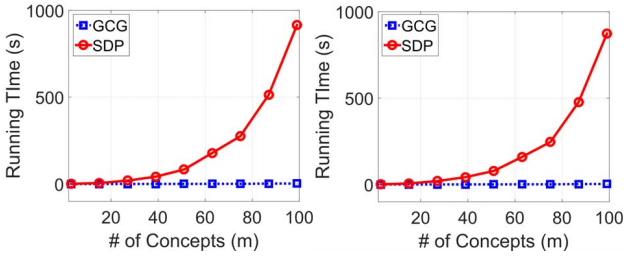
entire procedure in Algorithm 1. Following a similar argument as in [57], we can prove that Algorithm 1 converges globally to an ϵ-optimal solution of (12) in at most $O(1/\epsilon)$ steps. In practice, once we arrive at the true rank of the minimizer, the local solver (*e.g.* lbfgs) for the nonconvex problem (23) usually finds the solution at once. The per-step time complexity is $O(m^2)$ since the most time-consuming step is computing the rank-1 update in (21).

4. Experimental results

In this section we conduct thorough experiments to validate our warped spectral meta-learner for both the semantic event search and few-exemplar event detection tasks.

4.1. Speed comparison on synthetic data

We first verify the efficiency of the GCG Algorithm 1. We randomly generate m score vectors $\mathbf{s}_i \in \mathbb{R}^n, i = 1, \ldots, m$, and vary m from $m = 2$ to $m = 100$ (largest we were able to try). As can be seen from Figure 2, the running time of the naive SDP implementation (using YALMIP) increased sharply with the number of concepts. In comparison, the running time of our GCG implementation remains negligible (when achieving the same stopping criteria). It is clear that without our efficient GCG implementation, it is impossible to apply the (warped) spectral meta-learner to the large video datasets in the next section.



(a) λ = 1e-3 (b) λ = 1e-2

Figure 2: Efficiency comparison between GCG and SDP.

4.2. Experiment setup on real datasets

Datasets. We run experiments on three real datasets:

- MED14: The TRECVID MEDTest 2014 dataset [2] is collected by the NIST for all participants in the TRECVID competition. There are in total 20 events, whose description can be found in [2]. We use the official test split released by the NIST, and strictly follow its standard procedure [2]. In particular, we detect each event *separately*, treating each of them as a binary classification/ranking problem.
- MED13 [1]: Similar to MED14. Note that 10 of its 20 events overlap with those of MED14.
- CCV_{sub}: The official Columbia Consumer Video dataset [27] contains 9,317 videos in 20 semantic classes, including scenes like "beach," objects like "cat," and events like "baseball" and "parade." Since our goal is to *detect events*, we only use the 15 event categories.

We evaluate the performance using the mean Average Precision (mAP). Parameters of all compared algorithms are similarly tuned by grid search.

Concept detectors. 3,135 concept detectors are pre-trained using TRECVID SIN dataset (346 categories), Google sports (478 categories) [28, 24], UCF101 dataset (101 categories) [46, 24], YFCC dataset (609 categories) [3, 24] and DIY dataset (1601 categories) [56, 24]. We first extracted the improved dense trajectory features (including trajectory, HOG, HOF and MBH) using the code of [52] and encode them with the Fisher vector representation [41]. Following [52], we first reduce the dimension of each descriptor by a factor of 2 and then use 256 components to generate the Fisher vectors. Then, on top of the extracted low-level features, we trained the cascade SVM [19] for each concept detector. Using these concept detectors we obtain a 3,135-dimensional score vector for each video.

Competitors. We compare the following algorithms:

- Prim [20]: Primitive concepts, separately trained.
- Sel [35]: A subset of primitive concepts that are more informative for each event.
- Bi [43]: Bi-concepts discovered in [43].
- OR [20]: Boolean OR combinations of Prim concepts.
- Fu [20], Fu+: Boolean AND/OR combinations of Prim concepts, w/o concept refinement.
- Bor: The Borda rank aggregation in (2), with equal weights on the discovered semantic concepts.
- Bor+: Borda rank aggregation with equal weights on the refined semantic concepts.
- wBor: Borda rank aggregation with relevance weights on the discovered semantic concepts.
- wBor+: Borda rank aggregation with relevance weights on the refined semantic concepts.
- SML: Spectral meta-learner (15) on discovered semantic concepts.
- SML+: SML on refined semantic concepts.
- wSML: Warped SML (18) (with warping function (19)) on discovered semantic concepts.
- wSML+: Warped SML on refined semantic concepts.

The last eight methods are first proposed here. The refined concepts are subset of discovered semantic concepts, after dropping inaccurate, low predictive, and irrelevant ones.

Note that we did not compare with approaches that use multiple modalities of features, *e.g.* [25, 53], since we only considered the visual feature. In future work we plan to exploit speech and OCR information.

4.3. Semantic event search

We report the full experimental results on the TRECVID MEDTest 2014 dataset in Table 1 and also a summary on the MEDTest 2013 dataset and the CCV_{sub} dataset. We first consider the semantic event search setting where no labeled training video is available. As is clear from Table 1,

ID	Prim	Sel	Bi	OR	Fu	Fu+	Bor	Bor+	SML	SML+	wBor	wBor+	wSML	wSML+
	MEDTest 2014													
E021	2.12	2.98	2.64	3.89	3.97	4.43	2.67	3.46	4.29	5.14	3.12	4.64	5.37	6.48
E022	0.75	0.97	0.83	1.36	1.49	2.15	0.76	0.82	1.01	1.58	1.15	1.48	1.85	2.43
E023	33.86	36.94	35.23	39.18	40.87	42.62	35.65	37.22	38.19	41.73	38.68	41.78	43.26	50.55
E024	2.64	3.75	3.02	4.66	4.92	5.35	2.98	3.26	3.84	4.02	4.11	4.87	5.12	5.69
E025	0.54	0.76	0.62	0.97	1.39	1.87	0.52	0.75	0.92	1.06	0.84	1.01	1.26	1.43
E026	0.96	1.59	1.32	2.41	2.96	3.17	1.03	1.84	2.48	3.11	1.96	2.65	3.23	3.74
E027	11.21	13.64	12.48	15.93	16.26	18.56	12.52	12.73	13.96	15.62	15.12	16.47	18.63	20.56
E028	0.79	0.67	1.06	1.57	1.95	3.14	0.75	1.46	2.51	3.28	1.72	2.25	3.04	4.56
E029	8.43	10.68	12.21	14.01	14.85	16.52	9.64	10.25	11.93	13.48	13.19	14.75	16.69	18.84
E030	0.35	0.63	0.48	0.91	0.96	1.35	0.21	0.32	0.38	0.45	0.36	0.48	0.52	0.67
E031	32.78	53.19	45.87	69.52	69.66	72.59	54.29	61.82	65.75	70.43	67.49	72.64	76.45	82.86
E032	3.12	5.88	4.37	8.12	8.45	9.88	4.69	5.23	7.31	8.96	7.54	8.65	10.38	11.65
E033	15.25	20.19	18.54	22.14	22.23	25.07	17.66	18.71	19.49	22.04	21.53	23.26	25.64	28.93
E034	0.28	0.47	0.41	0.71	0.75	0.88	0.32	0.48	0.69	0.87	0.53	0.76	0.94	1.25
E035	9.26	13.28	11.09	16.53	16.68	19.26	12.74	14.95	16.28	19.49	15.82	18.65	20.78	25.39
E036	1.87	2.63	2.14	3.15	3.39	3.92	1.98	2.29	2.92	3.85	2.88	3.76	4.47	5.36
E037	2.16	4.52	3.81	6.84	6.88	7.26	3.26	4.19	5.33	6.41	5.42	6.83	7.45	8.68
E038	0.66	0.74	0.58	0.99	1.16	1.62	0.57	0.74	1.26	1.93	0.85	1.12	1.89	2.67
E039	0.36	0.57	0.42	0.69	0.77	0.97	0.45	0.58	0.83	1.02	0.64	0.85	1.26	1.73
E040	0.65	0.98	0.72	1.57	1.57	2.01	0.86	1.23	1.57	1.98	1.24	1.76	2.12	2.56
mean	6.40	9.55	7.89	10.76	11.05	12.13	8.18	9.12	10.05	11.33	10.21	11.44	12.52	14.32
	MEDTest 2013													
mean	7.07	7.94	6.92	9.45	9.88	10.62	6.86	7.61	8.79	10.08	8.43	9.96	11.64	13.46
	CCV_{sub}													
mean	19.05	19.40	20.25	21.16	21.89	22.52	21.19	22.23	22.66	23.42	23.08	23.87	24.71	25.59

Table 1: Experiment results for 0Ex event detection on MEDTest 2014, MEDTest 2013, and CCV_{Sub}. Mean average precision (mAP), in percentages, is used as the evaluation metric. Larger mAP indicates better performance.

the proposed methods (last eight columns) compare favorably against existing alternatives (first four columns), with a large margin obtained by the most sophisticated method wSML+ (14.32% vs the second best 10.76% achieved by OR) . The improvements are particularly impressive on some events, including *Dog Show (E23)*, *Rock Climbing (E27)*, *Beekeeping (E31)* and *Non-motorized vehicle repair (E33)*. By further looking into the discovered semantic concepts for these events, we find they all benefit greatly from relevant classifiers that are discriminative and reliable. For example, for the *Beekeeping* event, the performance of wSML+ significantly relied on concepts such as "apiary bee house" and "honeycomb", which turn out to be the most reliable concepts for *Beekeeping* in our concept vocabulary. Figure 3 shows the top 9 retrieved videos for the *Beekeeping* event. It is clear that videos retrieved by the proposed wSML+ are more accurate and visually coherent.

From Table 1 we make the following observations:

- Comparing the columns w/o the "+" suffix, we can see that concept refinement generally improves performance than naively using all concepts. This confirms the importance of using data-driven word embeddings to eliminate irrelevant and non-discriminative concepts.
- Comparing the Border and SML variants we verify the great benefit of using a principled method such as SML to combine classifiers. By taking into account the accuracy, albeit being estimated using unlabeled data, SML achieves better performance than simple majority voting.
- Comparing the columns w/o the "w" prefix, we observe that warping through the relevance significantly improves performance. This confirms the necessity to improve base classifiers, and illustrates the limitation of SML.

Similar conclusions can also be made from the results on MEDTest 2013 dataset and CCV_{sub} dataset.

4.4. Few-exemplar event detection

As mentioned in Section 3.6, our semantic event search framework can also be used for few-exemplar event detection: We simply combine the concept classifiers and the supervised classifier using the warped spectral meta-learner (with maximum weight for the latter). In this section, we demonstrate the benefit of this hybrid approach. Table 2 summarizes the mAP on both MEDTest 2014 and 2013, while Figure 4 compares the performance event-wise.

As a baseline (denoted as SVM in Table 2), we trained an SVM classifier using the improved dense trajectory (IDT) features [52] on 10 positive examples. Interestingly, this *supervised* classifier performed even slightly worse than our wSML+ which had no access to labeled data: mAP 13.92% vs 14.32% on MED14. This clearly demonstrates that a large pool of unsupervised but relevant concept classifiers, after proper correction of their accuracy, can outperform supervised classifiers trained with few positives. However, with more discriminative features such as convolutional neural networks (CNN), SVM with 10 positives outperformed wSML+: mAP 24.46% vs 14.32% on MED14.

Figure 4 gives a more detailed view of comparison: the supervised SVM (with IDT feature) is largely outperformed by our wSML+ on events E23, E24, E31, and E33, while the converse is observed on events E28, E30, E39, E40. In particular, on the event *Beekeeping* (E31), wSML+ improved

Figure 3: Top ranked videos for the event *Beekeeping*. From top to below: Sel (AP: 53.19), Bi (AP: 45.87), OR (AP: 69.52), and wSML+ (AP: 82.86). True/false labels (provided by NIST) are marked in the lower-right of each frame.

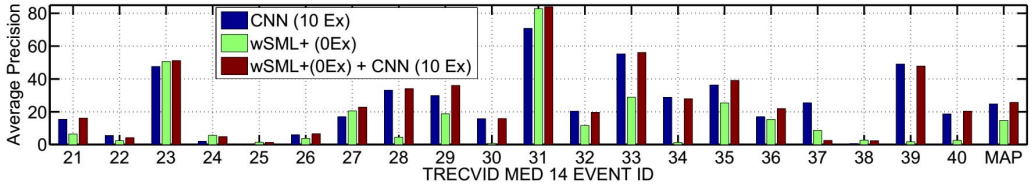


Figure 4: Performance comparison of CNN on MEDTest 2014 dataset, wSML+, and the hybrid of CNN and wSML+.

# of positives	Feature	Method	MED14	MED13
0	IDT	wSML	12.52	11.64
0	IDT	wSML+	14.32	13.46
10	IDT	SVM	13.92	18.08
10	IDT	wSML+ & SVM	16.98	19.65
10	CNN	SVM [45]	24.46	29.84
10	IDT & CNN	wSML+ & SVM	25.82	31.05

Table 2: Few-exemplar mAPs on MED14 and MED13.

IDT more than 2x (82.86% vs 33.9%). As mentioned before, this is because wSML+ significantly benefited from the presence of informative and reliable concepts such as "apiary bee house" and "honeycomb" on the particular *Beekeeping* event.

Finally we combine SVM with wSML+ as described before. This again significantly improves the performance from 13.92% to 16.98% on the MEDTest 2014 dataset. We also tried to combine with the state-of-the-art algorithm in [54], and increased its performance from 24.46% to 25.82%. As expected, the gain obtained from such simple hybrid diminishes when combining with more sophisticated methods. Overall, the results clearly demonstrate the utility of our framework even in the few-exemplar setting.

4.5. Annotated vs. unannotated data

We also compared the unsupervised SML approach with a supervised approach as follows. For each event we randomly select k labeled data from the MED14 training set, which are used to estimate the accuracy of the concept classifiers (*i.e.*, estimate p_i and n_i in (3)). Then we plug the estimates to the MLE (13) and obtain predictions on the test set. As can be seen from Figure 5, the unsupervised SML approach is advantageous when roughly 8 or less labeled training videos are used to estimate the concept accuracies. This experiment again demonstrates the utility of our method when the number of labeled training data is limited.

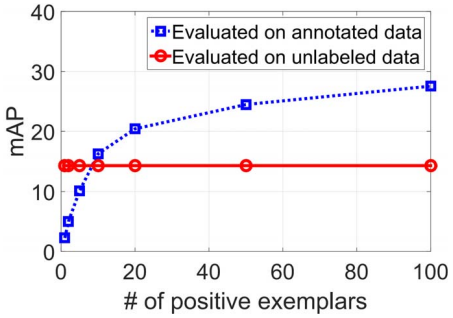


Figure 5: mAPs with increasing number of annotated pairs.

5. Conclusions

To address the challenging task of semantic event search and few-exemplar event detection, we proposed to leverage on concept classifiers collected from other sources. Data-driven word embedding models were used to seek the relevance of the concepts to the event of interest. To further account for the unreliability of the concept classifiers, we extended the recent spectral meta-learner that combines the classifiers based on a principled estimate of their accuracies using unlabeled data. Efficient implementations were provided and promising experimental results were obtained on three real video datasets. In the future we plan to incorporate temporal and spatial information [32, 48, 6] into our framework.

Acknowledgment

We thank the reviewers for their critical comments. This work was in part supported by the Data to Decisions Cooperative Research Centre www.d2dcrc.com.au, in part supported by NIH R01GM114311, in part supported by the ARC DECRA, and in part supported by the NSFC (U1509206).

References

[1] Trecvid MED 2013. http://nist.gov/itl/iad/mig/med13.cfm. 1, 6

[2] Trecvid MED 2014. http://nist.gov/itl/iad/mig/med14.cfm. 1, 3, 6

[3] The YFCC dataset. http://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67. 6

[4] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013. 3

[5] H. Bay, T. Tuytelaars, and L. J. V. Gool. SURF: speeded up robust features. In *ECCV*, 2006. 1

[6] S. Bhattacharya, M. M. Kalayeh, R. Sukthankar, and M. Shah. Recognition of complex events: Exploiting temporal dynamics between underlying concepts. In *CVPR*, 2014. 8

[7] Y. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010. 2

[8] L. Cao, Y. Mu, A. Natsev, S. Chang, G. Hua, and J. R. Smith. Scene aligned pooling for complex video recognition. In *ECCV*, pages 688–701, 2012. 2

[9] X. Chang, Z. Ma, Y. Yang, Z. Zeng, and A. G. Hauptmann. Bi-Level Semantic Representation Analysis for Multimedia Event Detection. *IEEE Transactions on Cybernetics*, 2016. 1, 2

[10] X. Chang, Y. Yang, A. G. Hauptmann, E. P. Xing, and Y. Yu. Semantic concept discovery for large-scale zero-shot event detection. In *IJCAI*, 2015. 1, 5

[11] X. Chang, Y. Yang, G. Long, C. Zhang, and A. G. Hauptmann. Dynamic concept composition for zero-example event detection. In *AAAI*, 2016. 1

[12] X. Chang, Y. Yang, E. P. Xing, and Y. Yu. Complex event detection using semantic saliency and nearly-isotonic SVM. In *ICML*, 2015. 1

[13] X. Chang, Y. Yu, Y. Yang, and A. G. Hauptmann. Searching persuasively: Joint event detection and evidence recounting with limited supervision. In *ACM MM*, 2015. 1

[14] J. Chen, Y. Cui, G. Ye, D. Liu, and S. Chang. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *ICMR*, 2014. 1

[15] Y. Cheng, Q. Fan, S. Pankanti, and A. N. Choudhary. Temporal sequence modeling for video event detection. In *CVPR*, 2014. 1

[16] J. Dalton, J. Allan, and P. Mirajkar. Zero-shot video retrieval using content and concepts. In *CIKM*, 2013. 1, 2

[17] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28(1):20–28, 1979. 2, 4

[18] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW*, 2001. 3

[19] H. P. Graf, E. Cosatto, L. Bottou, I. Durdanovic, and V. Vapnik. Parallel support vector machines: The cascade SVM. In *NIPS*, 2004. 6

[20] A. Habibian, T. Mensink, and C. G. M. Snoek. Composite concept discovery for zero-shot video event detection. In *ICMR*, 2014. 1, 2, 6

[21] A. Habibian, K. E. A. van de Sande, and C. G. M. Snoek. Recommendations for video event recognition using concept vocabularies. In *ICMR*, 2013. 1, 2

[22] A. Jaffe, B. Nadler, and Y. Kluger. Estimating the accuracies of multiple classifiers without labeled data. In *AISTATS*, 2015. 2, 4

[23] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *NIPS*, 2014. 2

[24] L. Jiang, D. Meng, S. Yu, Z. Lan, S. Shan, and A. G. Hauptmann. Self-paced learning with diversity. In *NIPS*, 2014. 6

[25] L. Jiang, T. Mitamura, S. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *ICMR*, page 297, 2014. 6

[26] Y. Jiang, S. Bhattacharya, S. Chang, and M. Shah. High-level event recognition in unconstrained videos. *IJMIR*, 2(2):73–101, 2013. 3

[27] Y. Jiang, G. Ye, S. Chang, D. P. W. Ellis, and A. C. Loui. Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In *ICMR*, 2011. 6

[28] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 6

[29] K. Lai, F. X. Yu, M. Chen, and S. Chang. Video event detection by inferring temporal instance labels. In *CVPR*, 2014. 1, 2

[30] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 3

[31] W. Li, Q. Yu, A. Divakaran, and N. Vasconcelos. Dynamic pooling for complex event recognition. In *ICCV*, 2013. 1, 2

[32] W. Li, Q. Yu, H. S. Sawhney, and N. Vasconcelos. Recognizing activities via bag of words for attribute dynamics. In *CVPR*, 2013. 8

[33] J. Liu, S. McCloskey, and Y. Liu. Local expert forest of score fusion for video event classification. In *ECCV*, 2012. 2

[34] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 1

[35] M. Mazloom, E. Gavves, K. E. A. van de Sande, and C. Snoek. Searching informative concept banks for video event detection. In *ICMR*, 2013. 6

[36] T. Mensink, E. Gavves, and C. G. M. Snoek. COSTA: co-occurrence statistics for zero-shot classification. In *CVPR*, 2014. 3

[37] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 1, 3

[38] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014. 3

[39] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009. 1

[40] F. Parisi, F. Strino, B. Nadler, and Y. Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111:1253–1258, 2014. 2, 4

[41] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, 2010. 2, 6

[42] E. A. Platanios, A. Blum, and T. Mitchell. Estimating accuracy from unlabeled data. In *UAI*, 2014. 2

[43] M. Rastegari, A. Diba, D. Parikh, and A. Farhadi. Multi-attribute queries: To merge or not to merge? In *CVPR*, 2013. 1, 2, 6

[44] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010. 2

[45] E. Roach and B. B. Lloyd. *Cognition and categorization*. 1978. 1

[46] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild, 2012. 6

[47] C. Sun and R. Nevatia. DISCOVER: discovering important segments for classification of video events and recounting. In *CVPR*, 2014. 1, 2

[48] K. D. Tang, F. Li, and D. Koller. Learning latent temporal structure for complex event detection. In *CVPR*, 2012. 8

[49] K. D. Tang, B. Yao, F. Li, and D. Koller. Combining the right features for complex event recognition. In *ICCV*, 2013. 2

[50] A. Vahdat, K. J. Cannons, G. Mori, S. Oh, and I. Kim. Compositional models for video event detection: A multiple kernel learning latent variable approach. In *ICCV*, 2013. 2

[51] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1582–1596, 2010. 1

[52] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013. 1, 6, 7

[53] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, 2014. 1, 2, 3, 6

[54] Z. Xu, Y. Yang, and A. G. Hauptmann. A discriminative CNN video representation for event detection. In *CVPR*, 2015. 1, 8

[55] G. Ye, D. Liu, I. Jhuo, and S. Chang. Robust late fusion with rank minimization. In *CVPR*, 2012. 3

[56] S. Yu, L. Jiang, and A. G. Hauptmann. Instructional videos for unsupervised harvesting and learning of action examples. In *ACM MM*, 2015. 6

[57] X. Zhang, Y. Yu, and D. Schuurmans. Accelerated training for matrix-norm regularization: A boosting approach. In *NIPS*, 2012. 5, 6