How Unlabeled Web Videos Help Complex Event Detection? *

Huan Liu[1], Qinghua Zheng[1], Minnan Luo[1], Dingwen Zhang[2], Xiaojun Chang[3]*, Cheng Deng[4]

[1]MOEKLINNS Lab, Department of Computer Science, Xi'an Jiaotong University, Shaanxi, China

[2]School of Automation, Northwestern Polytechnical University, Shaanxi, China

[3]School of Computer Science, Carnegie Mellon University, PA, USA

[4]School of Electronic Engineering, Xidian University, Shaanxi, China

Abstract

The lack of labeled exemplars is an important factor that makes the task of multimedia event detection (MED) complicated and challenging. Utilizing artificially picked and labeled external sources is an effective way to enhance the performance of MED. However, building these data usually requires professional human annotators, and the procedure is too time-consuming and costly to scale. In this paper, we propose a new robust dictionary learning framework for complex event detection, which is able to handle both labeled and easy-to-get unlabeled web videos by sharing the same dictionary. By employing the ℓ_q-norm based loss jointly with the structured sparsity based regularization, our model shows strong robustness against the substantial noisy and outlier videos from open source. We exploit an effective optimization algorithm to solve the proposed highly non-smooth and non-convex problem. Extensive experiment results over standard datasets of TRECVID MEDTest 2013 and TRECVID MEDTest 2014 demonstrate the effectiveness and superiority of the proposed framework on complex event detection.

1 Introduction

Multimedia event detection (MED) has become one of the most important visual content analysis tools as the rapid growth of web video-sharing services such as YouTube [Tamrakar *et al.*, 2012; Natarajan *et al.*, 2012; Habibian *et al.*, 2014; Chang *et al.*, 2015a]. This task aims to identify videos of a particular event of interest, *e.g.*, *making a cake* or *landing a fish*, which is the higher level semantic abstraction of long video clips consisting of multiple concepts [Yan *et al.*, 2015]. For example, an event like *landing a fish* can be described by multiple concepts, such as objects (*e.g.*, human, fish), actions (*e.g.*, standing, pulling) and scene (*e.g.*, beside a river or lake).

Detecting multimedia events from web videos is a complicated and challenging task due to the lack of labeled exemplars, especially the positive ones [Chang *et al.*, 2015b; 2017]. To enhance the performance of such complex event detection, traditional approaches rely on introducing some external sources that are relevant to the target events. Some researches assigned soft labels to related exemplars by assessing their "relatedness" with respect to the positive ones. With various concept selection strategies, Ye *et al.* and Ma *et al.* made use of high-level concept sources such as SIN dataset to facilitate the complex event detection [Ye *et al.*, 2015]. In [Duan *et al.*, 2012], event-related web videos were filtered out to help detect complex events. Generally, such external sources used by extant methods are artificially picked and labeled. However, building these external data usually requires professional human annotators, and the procedure is too time-consuming and costly to scale. On the contrary, there are plenty of low-cost unlabeled videos on the web, which might have significant information for the complex event detection. As a result, it would be beneficial to exploit a more flexible approach which is able to utilize the easy-to-get unlabeled web videos together with the limited labeled data.

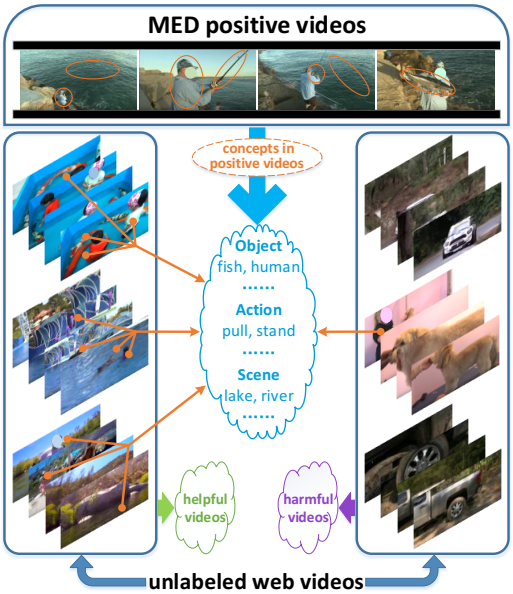

Figure 1: An example showing the influence of unlabeled web videos *w.r.t* the detection of the event *landing a fish*.

Open source videos, *i.e.*, unlabeled web videos without

*Corresponding author: Xiaojun Chang (cxj273@gmail.com)

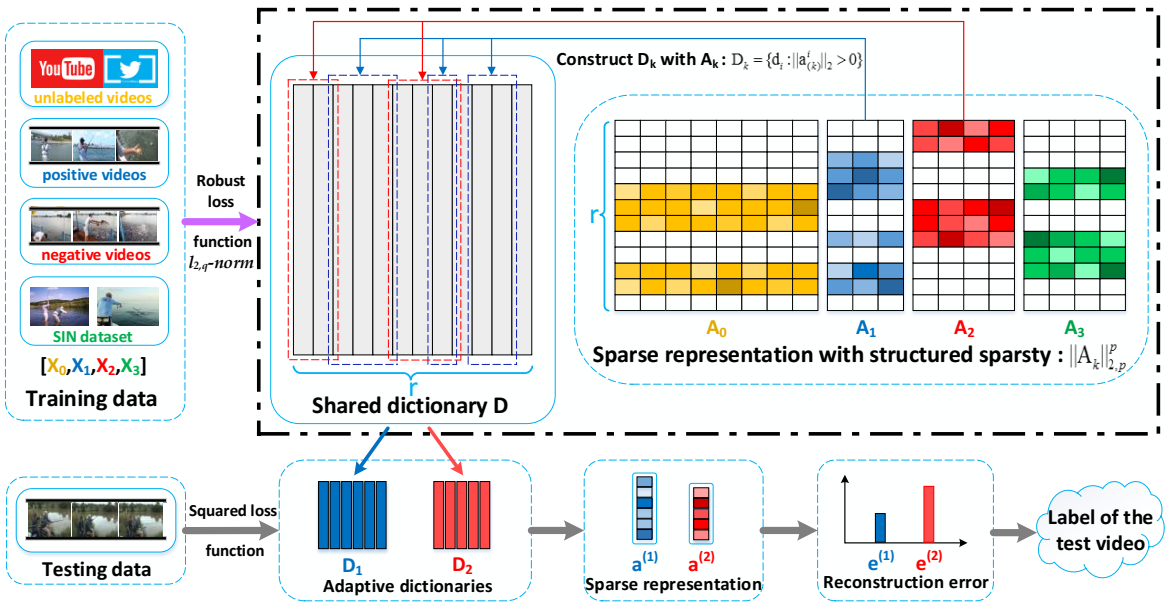

Figure 2: The framework for complex event detection. First, we represent both labeled and unlabeled data with a unified form for MED. Second, we design a robust dictionary learning framework to learn a shared dictionary D and structured sparsity based representation coefficients A_k. Next, we reduce the redundant information in dictionary D with A_k to generate adaptive dictionaries. Finally, we detect events from testing videos by using two adaptive dictionaries *w.r.t* positive and negative class.

artificial picking and annotating in this paper, are usually unstructured and do not follow any particular distribution [Chang *et al.*, 2015c]. Note that we call these videos unlabeled web videos because in this work they are used without any labels, even some weak labels such as the video's description. These characteristics make the unlabeled web videos easy to access. However, numerous videos from open source might have different impacts on the detection of a target event. To illustrate this point more clearly, we take the event *landing a fish* for an example (see Figure 1), where unlabeled web videos, *e.g.*, *playing with fishes* and *dolphin show*, share several concepts such as *fish* and *human* with the target event. These videos might be beneficial for the target event detection. However, there are also some unlabeled web videos, such as *motor racing* and *getting a vehicle unstuck*, which are completely irrelevant to the target event. Involving these web videos in the training stage might degrade the classifier. As a result, it is challenging to involve the easy-to-get web videos to enhance the performance of MED.

In light of this, there are mainly two issues to be considered with respect to introducing the open source videos into MED. The first issue is how to leverage the beneficial semantic information of unlabeled web videos to enhance the performance of complex event detection. The second issue is how to alleviate the side effect of unlabeled web videos, which might cause large uncertainties and potential threats. To address the both issues, we propose a new robust dictionary learning framework for complex event detection, which is able to handle both labeled and easy-to-get unlabeled web videos by sharing the same dictionary between them. Our model shows strong robustness against the substantial noisy and outlier videos from open source, as a result of employing the ℓ_q-norm $(0 < q < 2)$ based loss jointly with the structured sparsity based regularization. Figure 2 displays the overview of our framework for complex event detection with both labeled and unlabeled data. The objective is highly non-smooth and non-convex as using the ℓ_q-norm, therefore we present an effective alternating optimization algorithm to solve the proposed challenging problem.

We summarize our contributions as follows:

- Instead of picking and annotating web videos manually, we pioneer to employ the easy-to-get unlabeled web videos from open source to enhance the performance of complex event detection.
- To alleviate the side effect of varying unlabeled web videos, we propose a robust dictionary learning framework to handle labeled videos and unlabeled web videos simultaneously. This model shows strong robustness against the substantial noisy and outlier videos from open source.
- We conduct extensive experiments on the datasets of MED 13 and MED 14 for evaluation. The promising results demonstrate the effectiveness and superiority of the proposed method.

2 Related Work

2.1 MED with External Sources

External sources have been introduced to enhance the MED task, as the lack of labeled exemplars and the high cost of manual annotation. The intuitive way is to supplement the labeled samples directly. As high-level concept features are promising for event detection [Snoek *et al.*, 2006; Hauptmann *et al.*, 2007; Sadanand and Corso, 2012] and can be regarded as the bridge between training data and external

sources, several methods [Yan *et al.*, 2015] were presented to leverage external sources at concept level such as SIN dataset to facilitate event detection. These methods focused on transferring the knowledge from auxiliary data to the target events based on concept selection. However, there are still few researches on how to learn useful semantic information from open source rather than traditional dataset, which is labeled and well-structured but limited and difficult to extend, for the complex event detection. Several algorithms such as [Duan *et al.*, 2012] have made efforts to utilize web videos, which are filtered out with various strategies to ensure these external sources are related to the target event. As a result, these algorithms reduced the utilization rate of web videos and were time consuming for the real-world applications.

2.2 MED with Dictionary Learning

Dictionary Learning has the ability to represent the data with a set of relevant basis vectors. It has shown the state-of-the-art performance on computer vision analysis due to the intrinsic sparse property of visual data. Most of the researches were dedicated to image processing problem [Raina *et al.*, 2007; Mairal *et al.*, 2009b; Jiang *et al.*, 2015; Ramirez *et al.*, 2010], and few works focused on MED from videos. For example, Yan *et al.* proposed a novel event oriented dictionary representation in [Yan *et al.*, 2015] for each event to get more accurate and faster performance. While these works have made inspiring progress, the research on MED with dictionary learning is still in its infancy. Extant approaches are mainly based on traditional dictionary learning, which employs squared ℓ_2-norm as the reconstruction error. Therefore these methods are sensitive to noisy and outlier data. In the open source scenario, unlabeled web videos usually are unstructured and do not follow any particular distribution, hence noisy and outlier exemplars are inevitable by nature. As a result, robustness against noisy and outlier data needs to be taken into account in dictionary learning for complex event detection.

2.3 MED with Semi-Supervised Learning

Semi-supervised learning is a good way to handle the shortage of the labeled exemplars as its capability of utilizing both labeled and unlabeled data simultaneously. Recently, lots of researches [Chang *et al.*, 2014; Misra *et al.*, 2015; Liang *et al.*, 2015; Luo *et al.*, 2017] focused on the area of multimedia analysis such as feature selection and object detection with semi-supervised learning. For example, in [Chang *et al.*, 2014], Chang *et al.* proposed a novel convex semi-supervised multi-label feature selection algorithm for large-scale datasets. Inspired by the baby learning process, Liang *et al.* integrated prior knowledge modeling, exemplar learning and learning with video contexts for supervised object detection in [Liang *et al.*, 2015]. To avoid semantic drift when using semi-supervised learning [Misra *et al.*, 2015], Misra *et al.* presented an approach to constrain this process by combining multiple weak cues in videos and exploiting decorrelated errors by modeling data in multiple feature spaces. Though aforementioned work has shown the potential of semi-supervised learning, research on more challenging and complicated tasks such as MED is desired.

3 The Proposed Methodology

Considering the complexity of MED and the limited number of labeled exemplars for a specific event, TRECVID Semantic Indexing Task (SIN) dataset[1] is usually leveraged to enhance the performance of complex event detection. In this paper, we follow the concept selection strategy used in [Yan *et al.*, 2015]. For an event, first we calculate the similarity between concepts and this event by using both textual and visual methods, second we give equal weights to two different strategies to fuse the similarity and then select the Top 10 ranked concepts. As a result, we can automatically select videos which have semantic meaningful concepts for each MED event. Besides the SIN dataset which is labeled manually, a large number of videos are available without labels in practice. These videos might have significant information to enhance the accuracy of classification. However, few studies have been proposed to use these unlabeled videos for MED.

For a MED task, we suppose there are n_1 positive training exemplars of the event collected into matrix $X_1 = [\mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \cdots, \mathbf{x}_{n_1}^{(1)}] \in \mathbb{R}^{d\times n_1}$ and n_2 negative training exemplars of the event collected into matrix $X_2 = [\mathbf{x}_1^{(2)}, \mathbf{x}_2^{(2)}, \cdots, \mathbf{x}_{n_2}^{(2)}] \in \mathbb{R}^{d\times n_2}$. Let $X_3 = [\mathbf{x}_1^{(3)}, \mathbf{x}_2^{(3)}, \cdots, \mathbf{x}_{n_3}^{(3)}] \in \mathbb{R}^{d\times n_3}$ be n_3 semantic meaningful exemplars form SIN dataset, where the exemplars are collected from specified selected concepts for each event. $X_0 = [\mathbf{x}_1^{(0)}, \mathbf{x}_2^{(0)}, \cdots, \mathbf{x}_{n_0}^{(0)}] \in \mathbb{R}^{d\times n_0}$ be the matrix whose columns are the unlabeled exemplars. The goal of MED in this paper is to learn a shared dictionary across the exemplar matrices X_1, X_2, X_3, X_0 and the corresponding sparse representation of each exemplar, such that unlabeled videos from open source are involved in this task and the dictionary is able to predict labels for the unseen exemplars.

Let the shared dictionary $D = [\mathbf{d}_1, \cdots, \mathbf{d}_r] \in \mathbb{R}^{d\times r}$ consisting of r basis vectors $\mathbf{d}_j$ $(j = 1, 2, \cdots, r)$ and $\mathbf{a}_i^{(k)} = [a_{1i}^{(k)}, a_{2i}^{(k)}, \cdots, a_{ri}^{(k)}]^\top \in \mathbb{R}^r$ $(k = 0, 1, 2, 3; i = 1, 2, \cdots, n_k)$ be the sparse representation of exemplar $\mathbf{x}_i^{(k)} \in \mathbb{R}^d$ with respect to dictionary D. We collect all sparse representations of exemplar matrix X_k $(k = 0, 1, 2, 3)$ into a coefficient matrix $A_k = [\mathbf{a}_1^{(k)}, \cdots, \mathbf{a}_{n_k}^{(k)}] \in \mathbb{R}^{r\times n_k}$. In this paper, the optimal dictionary D and the sparse representation A_k $(k = 0, 1, 2, 3)$ are learned jointly by minimizing the following objective:

$$\begin{aligned} \min_{D,\{A_k\}_{k=0}^3} \quad & \sum_{k=0}^{3}\sum_{i=1}^{n_k} \|\mathbf{x}_i^{(k)} - D\mathbf{a}_i^{(k)}\|_2^q + \gamma \sum_{k=0}^{3} \|A_k\|_{2,p}^p \\ s.t. \quad & \mathbf{d}_j\mathbf{d}_j^\top \le 1,\ \forall j = 1, \ldots, r. \end{aligned} \tag{1}$$

where the constraints on the ℓ_2-norms of the basis vectors $\mathbf{d}_j$ $(j = 1, 2, \cdots, r)$ are utilized to avoid degenerate solutions [Wang *et al.*, 2013]. The ℓ_q-norm $(0 < q < 2)$ is used to measure the reconstruction errors over all exemplars. This strategy significantly mitigates the influence of a large number of outlier and noisy data which is derived very far from the true data distribution [Wang *et al.*, 2013].

[1] http://www-nlpir.nist.gov/projects/tv2013/tv2013.html#sin

Typically, the smaller the value of q is, the less impact the outlier (noisy) exemplars have. The $\ell_{2,p}$-norm regularization term $\|A_k\|_{2,p} = (\sum_{i=1}^{r}(\sum_{j=1}^{n_k}|a_{ij}^{(k)}|^2)^{\frac{p}{2}})^{\frac{1}{p}}$ penalizes all n_k entries in each row of representation coefficient matrix A_k which corresponds to one single basis vector of dictionary D as a whole. As a result, when $0 < p < 2$, structured sparsity is enforced on the coefficient matrix A_k $(k = 0, 1, 2, 3)$ such that the basis vector of dictionary corresponding to the non-zero row of coefficient matrix A_k is selected for succeeding data representation. Specifically, for the training exemplars $X_k \in \mathbb{R}^{d\times n_k}$, its dictionary is computed by

$$\mathcal{D}_k = \{\mathbf{d}_i : \|\mathbf{a}_{(k)}^i\|_2 > 0\} \tag{2}$$

where $\mathbf{a}_{(k)}^i$ $(i = 1, 2, \cdots, r)$ is the i-th row of spare representation matrix $A_k \in \mathbb{R}^{r\times n_k}$ $(k = 0, 1, 2, 3)$. For convenience, we construct the k-th dictionary for training exemplar matrix X_k as $D_k \in \mathbb{R}^{d\times|\mathcal{D}_k|}$ using all $\mathbf{d}_i \in \mathcal{D}_k$. Different from the method in [Wang *et al.*, 2013], which using unlabeled and labeled data, our model is able to leverage the easy-to-get unlabeled web videos, along with labeled MED data and event-related SIN dataset.

Given an unseen exemplar $\mathbf{x}$ and the learned dictionaries $D_k(1 \le k \le 2)$, we can compute the sparse representation $\mathbf{a}^{(k)}$ of $\mathbf{x}$ with regard to positive or negative class of the event, by solving the following LASSO problem:

$$\min_{\mathbf{a}^{(k)}} \|\mathbf{x} - D_k\mathbf{a}^{(k)}\|_2^2 + \gamma\|\mathbf{a}^{(k)}\|_1. \tag{3}$$

Based on the sparse representation, the reconstruction error of $\mathbf{x}$ with respect to positive or negative class is calculated by

$$\mathbf{e}^{(k)} = \|\mathbf{x} - D_k\mathbf{a}^{(k)}\|_2 \tag{4}$$

for $k = 1, 2$. Therefore, we can easily assign positive or negative label to this video $\mathbf{x}$ according to the result of sorting the reconstruction errors, *i.e.*,

$$l(\mathbf{x}) = \arg\min_k \mathbf{e}^{(k)}. \tag{5}$$

In summary, by sharing the same dictionary, our model is able to handle both labeled and unlabeled data with a unified form. Moreover, the ℓ_q-norm based loss and the structured sparsity based regularization ensure the strong robustness of the proposed model against the substantial noisy and outlier videos from open source. Finally, our model detect events from testing videos with positive and negative adaptive dictionaries, which are generated by the structured sparsity based representation coefficients.

4 Optimization Algorithm

Considering the non-smoothness and non-convexity of the objective function (1), in this section, we exploit an alternating optimization algorithm to solve the proposed challenging problem effectively. For a better representation, we first introduce some notations and rewrite the optimization problem (1) as its simpler form equivalently.

We denote $\mathcal{D} = \{D = [\mathbf{d}_1, \ldots, \mathbf{d}_r] \in \mathbb{R}^{d\times r} : \mathbf{d}_j\mathbf{d}_j^\top \le 1,\ \forall j = 1, \ldots, r\}$ as the feasible solution set of dictionary for optimization problem (1). Let $X = [X_0, X_1, X_2, X_3] \in \mathbb{R}^{d\times n}$ be the collection of all the training exemplars and $A = [A_0, A_1, A_2, A_3] \in \mathbb{R}^{r\times n}$ be the corresponding spare representation with respect to a specific dictionary $D \in \mathcal{D}$, where $n = \sum_{k=0}^{3} n_k$. By denoting $E = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n] = X - DA$ and $A_k = [\mathbf{a}_{(k)}^1, \mathbf{a}_{(k)}^2, \cdots, \mathbf{a}_{(k)}^r]^\top$ $(k = 0, 1, 2, 3)$, we define diagonal matrices $\Sigma = diag(\sigma_{11}, \sigma_{22}, \cdots, \sigma_{nn}) \in \mathbb{R}^{n\times n}$ and $\Lambda_k = diag(\lambda_{11}^{(k)}, \lambda_{22}^{(k)}, \cdots, \lambda_{rr}^{(k)}) \in \mathbb{R}^{r\times r}$ $(k = 0, 1, 2, 3)$, where[2]

$$\sigma_{ii} = \frac{1}{\frac{2}{q}\|\mathbf{e}_i\|_2^{2-q}}, \qquad \lambda_{jj}^{(k)} = \frac{1}{\frac{2}{p}\|\mathbf{a}_{(k)}^j\|_2^{2-p}} \tag{6}$$

for $i = 1, 2, \cdots, n$; $j = 1, 2, \cdots, r$ and $k = 0, 1, 2, 3$. In this sense, the optimization problem (1) is equivalent to the following problem:

$$\begin{aligned} &\min_{D,A} Tr\left((X - DA)\Sigma(X - DA)^\top\right) + \gamma\sum_{k=0}^{3} Tr(A_k^\top\Lambda_k A_k) \\ &s.t.\ \ D \in \mathcal{D} \end{aligned} \tag{7}$$

When variable A is fixed, the optimization problem (7) with respect to variable D turns to a traditional dictionary learning problem, *i.e.*,

$$\begin{aligned} &\min_{D\in\mathcal{D}} Tr\left((X - DA)\Sigma(X - DA)^\top\right) \\ \Leftrightarrow &\min_{D\in\mathcal{D}} \|(X\Sigma^{\frac{1}{2}} - DA\Sigma^{\frac{1}{2}})^\top\|_F^2. \end{aligned} \tag{8}$$

Considering the independence of the basis vectors, we update the dictionary D column by column through an efficient algorithm proposed in [Mairal *et al.*, 2009a].

When the shared dictionary D is fixed, the optimization problem (7) can be decomposed into four independent problems with respect to each sparse representation matrix A_k $(k = 0, 1, 2, 3)$, *i.e.*,

$$\min_{A_k} Tr\left((X_k - DA_k)\Sigma_k(X_k - DA_k)^\top\right) + \gamma Tr(A_k^\top\Lambda_k A_k) \tag{9}$$

where Σ_k is the k-th part of diagonal matrix Σ corresponding to training exemplar matrix X_k. By setting the derivative of the objective function above with respect to A_k to zero, we have $D^\top DA_k\Sigma_k - D^\top X_k\Sigma_k + \gamma\Lambda_k A_k = \mathbf{0}$. As a result, for each training exemplar $\mathbf{x}_i^{(k)}$ in X_k $(k = 0, 1, 2, 3; i = 1, 2, \cdots, n_k)$, we arrive at its spare representation given dictionary D by

$$\mathbf{a}_i^{(k)} = \sigma_{ii}^{(k)}(\sigma_{ii}^{(k)}D^\top D + \gamma\Lambda_k)^{-1}D^\top\mathbf{x}_i^{(k)}. \tag{10}$$

In summary, we describe the alternating algorithm for optimization problem (1) in Algorithm 1. Note that the main computational complexity of this algorithm lies in calculating

[2] σ_{ii} might be meaningless if $\frac{2}{q}\|\mathbf{e}_i\|_2^{2-q}$ is zero. Following the strategy used in [Nie *et al.*, 2010], we regularize σ_{ii} as $\frac{1}{\frac{2}{q}\|\mathbf{e}_i\|_2^{2-q}+\varsigma}$ since $\frac{1}{\frac{2}{q}\|\mathbf{e}_i\|_2^{2-q}+\varsigma} \to \frac{1}{\frac{2}{q}\|\mathbf{e}_i\|_2^{2-q}}$ when $\varsigma \to 0$. Similar strategy can be used for $\lambda_{jj}^{(k)}$.

Algorithm 1 Alternating algorithm for problem (1)

Input: Training data $X = [X_0, \ldots, X_K] \in \mathbb{R}^{d\times n}$, parameters γ, q and p.

Set $t = 0$, initialize $D \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times n}$ randomly, $\Sigma \in \mathbb{R}^{n\times n}$ and $\Lambda_k \in \mathbb{R}^{r\times r}$ $(0 \le k \le K)$ as identity matrices.

1: **repeat**
2: Update the dictionary D^{t+1} with algorithm proposed in [Mairal *et al.*, 2009a];
3: **for** $k = 0, \ldots, K$ **do**
4: Update each column of A_k^{t+1} with Eq. (10);
5: **end for**
6: Calculate the diagonal matrix Σ^{t+1} with the i-th diagonal element as $\sigma_{ii}^{t+1} = \frac{1}{\frac{2}{q}\|\mathbf{x}_i - D^{t+1}\mathbf{a}_i^{t+1}\|_2^{2-q}}$;
7: Calculate the diagonal matrix Λ_k^{t+1} $(0 \le k \le K)$ with the i-th diagonal element as $(\lambda_{ii}^{(k)})^{t+1} = \frac{1}{\frac{2}{p}\|(\mathbf{a}_{(k)}^i)^{t+1}\|_2^{2-p}}$;
8: $t = t + 1$
9: **until** Converge

Output: D and A.

the inverse of $r \times r$ matrix during updating sparse representation for each exemplar with complexity $\mathcal{O}(nr^3)$. To illustrate the effectiveness of the proposed algorithm theoretically, we first introduce a lemma in [Wang *et al.*, 2015], *i.e.*,

Lemma 1. *For any nonzero vectors $\mathbf{v}^{t+1}$ and $\mathbf{v}^t$, the following inequality holds for any $t = 0, 1, 2, \cdots$*

$$\|\mathbf{v}^{t+1}\|_2^p - \frac{\|\mathbf{v}^{t+1}\|_2^2}{\frac{2}{p}\|\mathbf{v}^t\|_2^{2-p}} \le \|\mathbf{v}^t\|_2^p - \frac{\|\mathbf{v}^t\|_2^2}{\frac{2}{p}\|\mathbf{v}^t\|_2^{2-p}} \qquad (11)$$

where $0 < p < 2$.

Based on Lemma 1, the convergence of the proposed alternating algorithm is illustrated by the following Theorem 1.

Theorem 1. *The alternating updating rules in Algorithm 1 monotonically decrease the objective function value of optimization problem* (1) *in each iteration.*

Proof. The detail proofs are skipped due to space limitation. It will be provided in the extended version of the paper. □

5 Experiment

5.1 Datasets

We evaluate on two large scale real-world datasets: the TRECVID MEDTest 2013 [3] and the TRECVID MEDTest 2014 [4], which are collected by the NIST for the TRECVID competition. Each dataset contains 20 complex events with 10 events in common. Specifically, the MEDTest 2013 dataset has events E006 to E015 and E021 to E030, while the MEDTest 2014 contains events E021 to E40. Please refer to the official website for the complete list of event names.

[3] http://nist.gov/itl/iad/mig/med13.cfm

[4] http://nist.gov/itl/iad/mig/med14.cfm

To enhance the performance of MED, SIN dataset is usually leveraged for its meaningful exemplars. It contains annotation for 346 semantic concepts from web videos. These concepts include objects, actions, scenes, attributes and non-visual concepts which are all the basic elements for an event, *e.g. kitchen, bus, boy*. We use the Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M) [Thomee *et al.*, 2016] as the unlabeled web videos in the experiments. YFCC100M is the largest public multimedia collection that has ever been released, comprising a total of 100 million media objects, of which approximately 99.2 million are photos and 0.8 million are videos. In this work, we only use these unlabeled videos (0.8m) to evaluate the proposed algorithm.

5.2 Experimental Setup

We extract state-of-the-art features from the videos. Specifically, Convolutional Neural Network (CNN) features are extracted. We extract the CNN features using the network architecture released by VGG. In our work, we extracted the features from the activation of the last pooling layer (pool5), as it has been demonstrated to be the best single feature in the literature. The VLAD encoding is applied after to form 65,536 dimensional representation. We cross-validated the regularization parameters in the range of $\{0.01, 0.1, 1, 10, 100\}$. We set $p = 0.8$ and $q = 1.2$ in our experiments to achieve the best performance. In order to keep our model (1) from being biased when the size of unlabeled data is greatly larger than labeled data, we follow the setting in [Amini and Gallinari, 2002], which showed that 20% of labeled data in the training set could improve the performance significantly. Average Precision (AP) and Mean AP (mAP) are used as the evaluation metrics. Higher value indicates better performance.

5.3 Comparison Methods

The proposed algorithm is compared with these baselines:

- Support Vector Machine (SVM): SVM has been widely used by several research groups for MED competition and has shown its effectiveness. Hence, we use it as a baseline.
- Multiple Kernel Transfer Learning (MKTL) [Luo *et al.*, 2011]: This algorithm aims to incorporate prior features into a multiple kernel learning framework.
- Dirty Model Multi-Task Learning (DMMTL) [Jalali *et al.*, 2010]: This is a state-of-the-art multi-task learning algorithm imposing ℓ_1/ℓ_p-norm regularizations.
- Multiple Kernel Learning Latent Variable Approach (MKLLVA) [Vahdat *et al.*, 2013]: It is a multiple kernel learning latent variable approach for complex event detection.
- Event Oriented Dictionary Learning (EODL) [Yan *et al.*, 2015]: EODL leverages training samples of selected concepts from the SIN dataset into a jointly supervised multi-task dictionary learning framework.
- Class-Specific Sparse Multiple Kernel Learning (CSSMKL) [Liu *et al.*, 2016]: This is a class-specific sparse multiple kernel learning algorithm for spectral–spatial hyperspectral image classification.

5.4 Experimental Results Analysis

To demonstrate the effectiveness of our proposed robust dictionary learning framework with unlabeled web videos, we

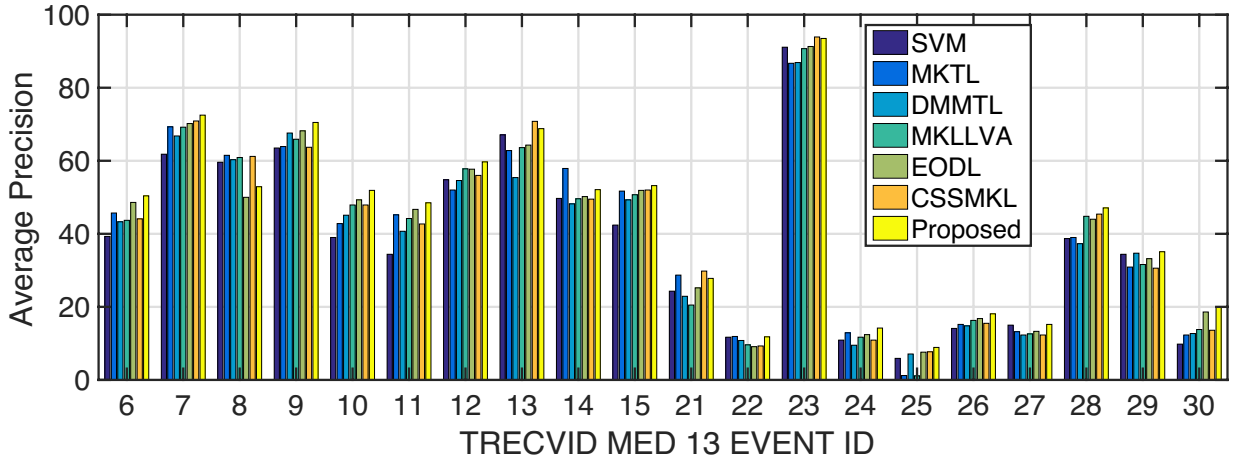

Figure 3: Comparison of different methods of AP performance for event on MEDTest 2013 (100Ex).

Table 1: Mean average precision (mAP) comparison of different methods on MEDTest 2013 and MEDTest 2014.

	MED13		MED14	
	100Ex	10Ex	100Ex	10Ex
SVM	38.2	25.0	33.9	22.8
MKTL[Luo *et al.*, 2011]	39.8	26.3	35.2	24.1
DMMTL[Jalali *et al.*, 2010]	39.1	25.8	34.4	23.7
MKLLVA[Vahdat *et al.*, 2013]	40.4	26.8	35.9	24.8
EODL[Yan *et al.*, 2015]	42.5	27.5	37.2	25.6
CSSMKL[Liu *et al.*, 2016]	41.8	27.7	36.8	25.3
The Proposed model	**44.2**	**29.5**	**38.6**	**27.7**

evaluate all comparison methods with 10 and 100 positive exemplars from test datasets, respectively. Due to the space limitation, we only present in Figure 3 the performance of AP with respect to each event of TRECVID MED 13. We observe from the experimental results that: (1) Since both the proposed model and EODL are based on dictionary learning, they generally perform as the best competitive classifiers, while other algorithms have varying degree of success on different events. (2) Thanks to the unlabeled web videos introduced in complex event detection, our algorithm achieves the best or second-best performance for 19 out of 20 events, indicating the positive function of unlabeled web videos to some extent. Specifically, for almost all of the events, such as "E006: Birthday party" and "E010: Grooming an animal", there are plenty of related videos in YFCC100M which play an effective role in promoting the performance of complex event detection; Instead, only a few unlabeled videos from YFCC101M are relevant to the event "E008: Flash mob gathering". These substantial unlabeled web videos are served as noises and outliers for complex event detection. As a result, it is reasonable that our algorithms performs poorer than EODL in terms of "E008: Flash mob gathering."

For a fair comparison, we further report in Table 1 the values of mAP over 20 events of TRECVID MEDTest 2013 and 2014, respectively. The experimental results indicate that: (1) For all of the competitors, more positive exemplars, *i.e.*, 100Ex vs. 10Ex, are beneficial to enhance the performance of complex event detection remarkably. This phenomenon gears to the fact that more labeled data is crucial for MED tasks; (2) Our model consistently outperforms other state-of-the-art methods with respect to both 100Ex and 10Ex. Specifically, our model achieves a 15.8% on average improvement in terms of mAP comparing with the SVM that is widely used in MED competition. This result indicates that the proposed robust dictionary learning framework has strong robustness against substantial noisy and outlier videos from open source.

6 Conclusion

In this paper, we have carried out the first exploration about utilizing the easy-to-get unlabeled web videos to enhance the performance of complex event detection. We proposed a new robust dictionary learning framework to handle both labeled and unlabeled data by addressing two main challenges caused by open source videos, *i.e.*, how to leverage their beneficial semantic information and how to alleviate their side effect. Moreover, we present an effective alternating optimization algorithm to solve the proposed highly non-smooth and non-convex problem. Finally, extensive experiment results on several large scale datasets demonstrate the effectiveness of the proposed method. In the future, we intend to focus on the dictionary learning over multi-view features for MED.

Acknowledgments

This work is was supported in part by "The Fundamental Theory and Applications of Big Data with Knowledge Engineering" under the National Key Research and Development Program of China with grant Nos 2016YFB1000903; Ministry of Education Innovation Research Team No. IRT13035; Project of China Knowledge Centre for Engineering Science and Technology; National Science Foundation of China under Grant Nos. 61502377.

References

[Amini and Gallinari, 2002] Massih-Reza Amini and Patrick Gallinari. The use of unlabeled data to improve supervised learning for text summarization. In *SIGIR*, 2002.

[Chang *et al.*, 2014] Xiaojun Chang, Feiping Nie, Yi Yang, and Heng Huang. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, 2014.

[Chang *et al.*, 2015a] Xiaojun Chang, Yi Yang, Alexander Hauptmann, Eric P Xing, and Yao-Liang Yu. Semantic concept discovery for large-scale zero-shot event detection. In *IJCAI*, 2015.

[Chang *et al.*, 2015b] Xiaojun Chang, Yi Yang, Eric P. Xing, and Yaoliang Yu. Complex event detection using semantic saliency and nearly-isotonic SVM. In *ICML*, 2015.

[Chang *et al.*, 2015c] Xiaojun Chang, Yao-Liang Yu, Yi Yang, and Alexander G Hauptmann. Searching persuasively: Joint event detection and evidence recounting with limited supervision. In *ACM MM*, 2015.

[Chang *et al.*, 2017] Xiaojun Chang, Zhigang Ma, Yi Yang, Zhiqiang Zeng, and Alexander G. Hauptmann. Bi-level semantic representation analysis for multimedia event detection. *IEEE Trans. Cybernetics*, 47(5):1180–1197, 2017.

[Duan *et al.*, 2012] Lixin Duan, Dong Xu, Ivor Wai-Hung Tsang, and Jiebo Luo. Visual event recognition in videos by learning from web data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1667–1680, 2012.

[Habibian *et al.*, 2014] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACM MM*, 2014.

[Hauptmann *et al.*, 2007] Alexander Hauptmann, Rong Yan, Wei-Hao Lin, Michael Christel, and Howard Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *IEEE Trans. Multimedia*, 9(5):958–966, 2007.

[Jalali *et al.*, 2010] Ali Jalali, Pradeep Ravikumar, Sujay Sanghavi, and Chao Ruan. A dirty model for multi-task learning. In *NIPS*, 2010.

[Jiang *et al.*, 2015] Wenhao Jiang, Feiping Nie, and Heng Huang. Robust dictionary learning with capped ℓ_1-norm. In *IJCAI*, 2015.

[Liang *et al.*, 2015] Xiaodan Liang, Si Liu, Yunchao Wei, Luoqi Liu, Liang Lin, and Shuicheng Yan. Towards computational baby learning: A weakly-supervised approach for object detection. In *ICCV*, 2015.

[Liu *et al.*, 2016] Tianzhu Liu, Yanfeng Gu, Xiuping Jia, Jón Atli Benediktsson, and Jocelyn Chanussot. Class-specific sparse multiple kernel learning for spectral–spatial hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.*, 54(12):7351–7365, 2016.

[Luo *et al.*, 2011] Jie Luo, Tatiana Tommasi, and Barbara Caputo. Multiclass transfer learning from unconstrained priors. In *ICCV*, 2011.

[Luo *et al.*, 2017] Minnan Luo, Xiaojun Chang, Liqiang Nie, Yi Yang, Alexander G Hauptmann, and Qinghua Zheng. An adaptive semisupervised feature analysis for video semantic recognition. *IEEE Trans. Cybern.*, PP(99):1–13, 2017.

[Mairal *et al.*, 2009a] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *ICML*, 2009.

[Mairal *et al.*, 2009b] Julien Mairal, Jean Ponce, Guillermo Sapiro, Andrew Zisserman, and Francis R Bach. Supervised dictionary learning. In *NIPS*, 2009.

[Misra *et al.*, 2015] Ishan Misra, Abhinav Shrivastava, and Martial Hebert. Watch and learn: Semi-supervised learning for object detectors from video. In *CVPR*, 2015.

[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.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, 2010.

[Raina *et al.*, 2007] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. Self-taught learning: transfer learning from unlabeled data. In *ICML*, 2007.

[Ramirez *et al.*, 2010] Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *CVPR*, 2010.

[Sadanand and Corso, 2012] Sreemanananth Sadanand and Jason J Corso. Action bank: A high-level representation of activity in video. In *CVPR*, 2012.

[Snoek *et al.*, 2006] Cees GM Snoek, Marcel Worring, Jan C Van Gemert, Jan-Mark Geusebroek, and Arnold WM Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *ACM MM*, 2006.

[Tamrakar *et al.*, 2012] Amir Tamrakar, Saad Ali, Qian Yu, Jingen Liu, Omar Javed, Ajay Divakaran, Hui Cheng, and Harpreet Sawhney. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *CVPR*, 2012.

[Thomee *et al.*, 2016] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, 2016.

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

[Wang *et al.*, 2013] Hua Wang, Feiping Nie, Weidong Cai, and Heng Huang. Semi-supervised robust dictionary learning via efficient $\ell_{2,0^+}$-norms minimization. In *ICCV*, 2013.

[Wang *et al.*, 2015] Hua Wang, Feiping Nie, and Heng Huang. Learning robust locality preserving projection via p-order minimization. In *AAAI*, 2015.

[Yan *et al.*, 2015] Yan Yan, Yi Yang, Deyu Meng, Gaowen Liu, Wei Tong, Alexander G Hauptmann, and Nicu Sebe. Event oriented dictionary learning for complex event detection. *IEEE Trans. Image Process.*, 24(6):1867–1878, 2015.

[Ye *et al.*, 2015] Guangnan Ye, Yitong Li, Hongliang Xu, Dong Liu, and Shih-Fu Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *ACM MM*, 2015.