Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Complex Event Detection via Event Oriented Dictionary Learning

Yan Yan,[1,*] **Yi Yang,**[2] **Haoquan Shen,**[2]
Deyu Meng,[3] **Gaowen Liu,**[1] **Alex Hauptmann,**[4] **Nicu Sebe**[1]
[1]University of Trento, Italy [2]University of Technology Sydney, Australia
[3]Xi'an Jiao Tong University, China [4]Carnegie Mellon University, USA
yan@disi.unitn.it, yee.i.yang@gmail.com, shenhaoquan@gmail.com,
dymeng@mail.xjtu.edu.cn, gaowen.liu@unitn.it, alex@cs.cmu.edu, sebe@disi.unitn.it

Abstract

Complex event detection is a retrieval task with the goal of finding videos of a particular event in a large-scale *unconstrained* internet video archive, given example videos and text descriptions. Nowadays, different multimodal fusion schemes of low-level and high-level features are extensively investigated and evaluated for the complex event detection task. However, how to effectively select the high-level semantic meaningful concepts from a large pool to assist complex event detection is rarely studied in the literature. In this paper, we propose two novel strategies to *automatically* select semantic meaningful concepts for the event detection task based on both the events-kit text descriptions and the concepts high-level feature descriptions. Moreover, we introduce a novel event oriented dictionary representation based on the selected semantic concepts. Towards this goal, we leverage training samples of selected concepts from the Semantic Indexing (SIN) dataset with a pool of 346 concepts, into a novel supervised multi-task dictionary learning framework. Extensive experimental results on TRECVID Multimedia Event Detection (MED) dataset demonstrate the efficacy of our proposed method.

Introduction

Complex event detection in unconstrained videos has received much attention in the research community recently (Tamrakar et al. 2012; Ma et al. 2013; Natarajan et al. 2012). It is a retrieval task with the goal of detecting videos of a particular event in a large-scale internet video archive, given an event-kit. An event-kit consists of example videos and text descriptions of the event. Unlike traditional action recognition of atomic actions, such as 'walking' or 'jumping' from videos, complex event detection aims to detect more complex events such as 'Birthday party', 'Changing a vehicle tire', *etc*.

An *event* is a higher level semantic abstraction of video sequences than a *concept* and consists of many *concepts*. For example, a 'Birthday party' event can be described by multiple concepts, such as objects (*e.g.*, boy, cake), actions (*e.g.*, talking, walking) and scene (*e.g.*, at home, in a restaurant). A concept can be detected in a shorter video sequence or even in a single frame but an event is usually contained in a longer video clip.

Traditional approaches for complex event detection rely on fusing multiple low-level features classification outputs (Tamrakar et al. 2012), *i.e.* SIFT, STIP, MOSIFT (Chen and Hauptmann 2009). Recently, representing videos using high-level features, such as concept detectors (Snoek and Smeulders 2010), appears promising for the complex event detection task. However, the state-of-the-art concept detector based approaches (Jiang, Hauptmann, and Xiang 2012; Snoek and Smeulders 2010; Ma et al. 2013; Sun and Nevatia 2013) for complex event detection have not considered which concepts should be included in the training concept list. This always conducts the redundancy of concepts (Ma et al. 2013; Sun and Nevatia 2013) in the concept list for the vocabulary construction. For example, it is highly impossible for some concepts to help detect certain event, *e.g.* 'cows', 'football' are not helpful to detect events like 'Landing a fish' or 'Working on a sewing project'. Therefore, removing the uncorrelated concepts from the vocabulary construction inclines to eliminate such redundancy and boost the complex event detection performance.

Intuitively, it is highly expected to be more accurate and faster for complex event detection when we build specific dictionary representation for each event. In this paper, we investigate how to learn a concept-driven event oriented representation for complex event detection. There are mainly two-fold important issues to be considered to accomplish this goal. The first issue is which concepts should be included in the vocabulary construction of the learning framework. Since we want to learn an event oriented dictionary representation, how to properly select qualified concepts for each event in the learning framework is the key issue. This raises the problem of how to optimally select necessary and meaningful concepts from a large pool of concepts for each event. The second issue is how can we design an effective dictionary learning framework to seamlessly learn the common knowledge from both the low-level features and the high-level concept features.

To facilitate reading, we first describe some abbreviations used in the paper. SIN stands for Sematic Indexing which is a dataset[1] containing 346 different categories (concepts)

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

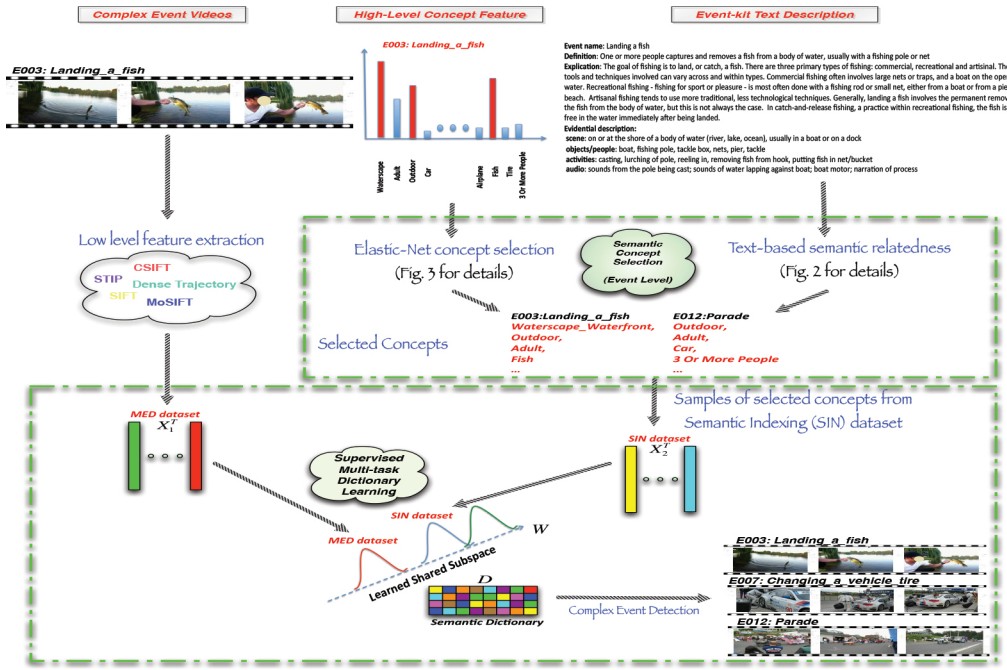

Figure 1: Illustration of our event oriented dictionary learning framework. *Top*: Different complex video events with their high-level concept feature and the corresponding event-kit text description. *Middle left*: Different types of low-level features extraction for events. *Middle right*: Event specific concept pool construction based on (i) the high-level concept feature descriptions using Elastic-Net concept selection, (ii) the MED events-kit text descriptions using linguistic knowledge. *Bottom*: Supervised multi-task dictionary learning. (Figure is best viewed in color and under zoom).

of images, such as car, adult, *etc*. SIN-MED stands for the high-level concept features using the SIN concept list representing each MED video by a 346-dimensional feature (each dimension represents a concept).

The overview of our framework is shown in Fig.1. Firstly, we design two novel methods to *automatically* select semantic meaningful concepts for each MED event based on both MED events-kit text descriptions and SIN-MED high-level concept feature representations. Then we leverage training samples of selected concepts from the SIN dataset into a jointly supervised multi-task dictionary learning framework. An event specific semantic meaningful dictionary is learned through embedding the feature representation of original datasets (both MED dataset and SIN dataset) into a hidden shared subspace. We add label information in the learning framework to facilitate the event oriented dictionary learning process. Therefore, the learned sparse codes achieve intrinsic discriminative information and naturally lead to the effectiveness of complex event detection.

To summarize, the contributions of this paper are as follows: (i) We propose two novel approaches of concept selection strategies and present one of the first works to make a comprehensive evaluation for automatic concept selection for event detection. (ii) We are the first to propose the event oriented dictionary learning for event detection. (iii) We firstly construct a supervised multi-task dictionary learning framework which is capable of learning an event oriented dictionary via leveraging information from selected semantic concepts. (iv) The proposed learning framework is a generic one which can be easily generalized into other computer vision and pattern recognition problems.

Related Work

Complex Event Detection. With the success of event detection in structured scenarios, complex event detection from general *unconstrained* videos, such as those obtained from internet video sharing web sites like YouTube, has been receiving increasing attention in recent years. Tamrakar *et al.* (Tamrakar et al. 2012) evaluated different low-level appearance as well as spatio-temporal features, appropriately quantized and aggregated them into Bag-of-Words (BoW) descriptors for complex event detection. Natarajan *et al.* (Natarajan et al. 2012) evaluated a large set of low-level audio and visual features as well as high-level information from object detection, speech and video text OCR for complex event detection. They combined multiple features using a multi-stage feature fusion strategy with feature level early fusion using multiple kernel learning, score level fusion using Bayesian model combination and weighted average fusion using video specific weights. Vahdat *et al.* (Vahdat et al. 2013) presented a compositional model for complex event detection that leveraged a novel multiple kernel learning algorithm to incorporate structured latent variables. However, to the best of our knowledge, there are still few research works on how to automatically select useful high-level concepts for the complex event detection.

Dictionary Learning. Dictionary Learning has been verified to be able to find succinct representations of stimuli and model data vectors as a linear combination of a few elements from a dictionary. Dictionary learning has been successfully applied to a variety of problems in computer vision analysis recently, *e.g.* image classification (Yang et al. 2009), image denoising (Elad and Aharon 2006) and image segmentation (Mairal et al. 2008). Different optimization algorithms (Aharon, Elad, and Bruckstein 2006; Lee et al. 2006) have also been proposed to solve dictionary learning problems. However, so far as we know, there is no research work on how to learn the dictionary representation at the event level for event detection and there is no research work on how to simultaneously leverage the semantic information to learn an event oriented dictionary.

Multi-task Learning. Multi-task learning (Argyriou, Evgeniou, and Pontil 2007) methods aim to simultaneously learn classification/regression models for a set of related tasks. This typically leads to better models as compared to separately consider each task without accounting for task relationships. The goal of multi-task learning is to improve the performance of learning algorithms by learning classifiers for multiple tasks jointly. This works particularly well if these tasks have some commonality while are all slightly under-sampled. The effectiveness of multi-task Learning has been demonstrated in several applications in computer vision, such as headpose classification (Yan et al. 2013a) and action recognition (Yan et al. 2013b). However, there is few work on multi-task learning used for dictionary learning problem. Maurer *et al.* (Maurer, Pontil, and Paredes 2013) only provides theoretical bounds to evaluate the generalization error of dictionary learning for multi-task learning and transfer learning.

Building Event Specific Concept Pool

The concepts, which are related to objects, actions, scenes, attributes, *etc.* are usually basic elements for the description of an event. There are usually a large pool of concept detectors exsited for event descriptions since the availability of large labeled training collections such as ImageNet (Berg et al. 2011) and TRECVID (Smeaton, Over, and Kraaij 2006). However, selecting important concepts are the key issues for concept vocabulary construction. For example, the event 'Landing a fish' is composed of the most important concepts such as 'adult', 'waterscape', 'outdoor' and 'fish'. If we can get these concepts intrinsically related to the interested event, the concept redundancy problem tends to be ameliorated and the complex event detection performance inclines to be further boosted. In order to select useful concepts for the specific event, we propose two novel concept selection strategies in this section, which are (i) Text-based semantic relatedness from linguistic knowledge of MED event-kit text description and (ii) Elastic-Net feature selection from visual high-level representation.

Linguistic: Text-based Semantic Relatedness

[2]http://www.nist.gov/itl/iad/mig/med12.cfm

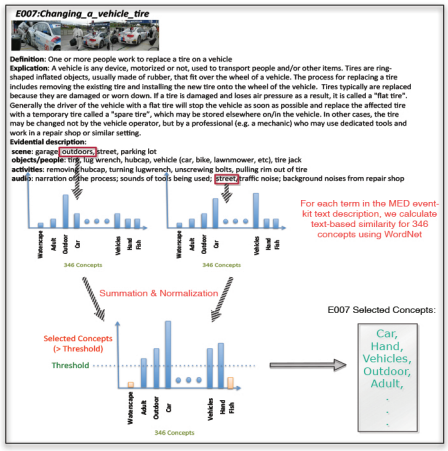

Figure 2: Linguistic-based concept selection strategy with an example of 'E007: Changing a vehicle tire' in MED event-kit text description and a corresponding example video provided by NIST (Figure is best viewed under zoom).

The most widely used resources in Natural Language Processing (NLP) to calulate the semantic relateness of concepts are WordNet (Fellbaum 1998) and Wikipedia (Strube and Ponzetto 2006). There are detailed event-kit text descriptions for each MED event provided by NIST[2]. In this paper, we explore the semantic similarity between each term in the event-kit text description and SIN 346 visual concept names based on WordNet. Fig.2 shows an example of event-kit text description for 'Changing a vehicle tire'.

As illustrated in Fig.2, we calculate the similarity between each term in event-kit text descriptions and the SIN 346 visual concept names based on the similarity measurement proposed in (Lin 1998). This measurement defines the similarity of two words w_{1i} and w_{2j} as:

$$sim(w_{1i}, w_{2j}) = \frac{2\,\pi(l_{cs})}{\pi(w_{1i}) + \pi(w_{2j})}$$

where $w_{1i} \in \{\text{event-kit text descriptions}\}$, $i = 1, \ldots, N^{event\_kit}$ and $w_{2j} \in \{\text{SIN visual concept names}\}$, $j = 1, \ldots, 346$. l_{cs} denotes the lowest common subsumer of two words in the WordNet hierarchy. π denotes the information content of a word and is computed as $\pi(w) = log\, p(w)$, where $p(w)$ is the probability of encountering an instance of w in a corpus. The probability $p(w)$ can be estimated from the relative corpus frequency of w and the probabilities of all words that w subsumes (Resnik 1995). In this way, we expect to properly capture the semantic similarity between subjects (*e.g.* human, crowd) and objects (*e.g.* animal, vehicle) based on the WordNet hierarchy. Finally, we construct a 346-dimensional event-level feature vector representation for each event (each dimension corresponds to an SIN visual concept name) using the MED event-kit text description from linguistic knowledge. A threshold is set ($thr = 0.5$ in our experiments) to select useful concepts into our final semantic concept list.

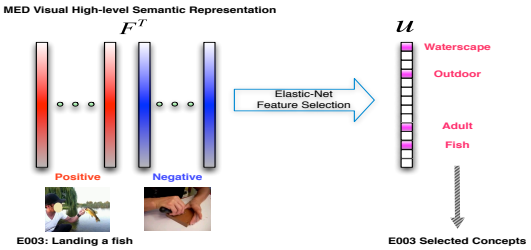

Figure 3: Visual high-level semantic representation with Elastic-Net concept selection.

Visual High-level Representation: Elastic-Net Concept Selection

Concept detectors provide a high-level semantic representation for videos with complicated contents, which incline to benefit for developing powerful retrieval or filtering systems for consumer media (Snoek and Smeulders 2010). In our case, we firstly use the SIN dataset to train 346 semantic concept models. In the meanwhile, we adopt (Luo, Papin, and Costello 2009) to extract keyframes from the MED dataset. The trained 346 semantic concept models are used to predict the 346 semantic concepts existing in the keyframes of MED videos. Once we have the prediction score of each concept on each keyframe, the keyframe can be represented as a 346-dimensional SIN-MED feature indicating the determined concept probabilities. Finally, the video-level SIN-MED feature is computed as the average of keyframe-level SIN-MED feature.

To select the useful concepts for each specific event, we adopt the Elastic-Net (Zou and Hastie 2005) concept selection as illustrated in Fig.3, given the intuition that the learner generally would like to choose the most representative SIN-MED feature dimensions (concepts) to differentiate events. Elastic-Net is formulated as follows:

$$\min_{\mathbf{u}} \|\mathbf{l} - \mathbf{F}\mathbf{u}\|^2 + \alpha_1 \|\mathbf{u}\|_1 + \alpha_2 \|\mathbf{u}\|^2$$

where $\mathbf{l} = \{0,1\}^n \in I\!R^n$ indicates the event labels, $\mathbf{F} \in I\!R^{n \times b}$ is the SIN-MED feature matrix (n is the number of samples and b is the SIN-MED feature dimension) and $\mathbf{u} \in I\!R^b$ is the parameter to be optimized. Each dimension of $\mathbf{u}$ corresponds to one semantic concept if $\mathbf{F}$ is the high-level SIN-MED feature. α_1 and α_2 are the regularization parameters. We use Elastic-Net instead of LASSO due to the high correlation between concepts in the SIN concept lists. While LASSO (when $\alpha_2 = 0$) tends to select only a small number of variables from a group and ignore the others, Elastic-Net is capable of automatically taking such correlation information into account through adding a quadratic term $\|\mathbf{u}\|^2$ to the penalty. We can adjust the value of α_1 value to control the sparsity degree, *i.e.*, how many semantic concepts are selected in our problem. The concepts to be selected are the corresponding dimensions with non-zero vaules of $\mathbf{u}$.

To sum up, we combine the semantic concepts selected from both human linguistic as described in section and visual high-level semantic representation as described in section to form the final list of selected concepts for each MED event.

Event Oriented Dictionary Learning

After we select semantic meaningful concepts for each event, we can leverage training samples of selected concepts from the SIN dataset into a supervised multi-task dictionary learning framework. In this section, we investigate how to learn an event oriented dictionary representation. To accomplish this goal, we firstly propose our multi-task ditionary learning framework and then introduce its supervised setting.

Multi-task Dictionary Learning

Given K tasks (*e.g.* $K = 2$ in our case, one task is the MED dataset and the other task is the subset of SIN dataset where samples are collected from specified selected concepts for each event), each task consists of data samples denoted by $\mathbf{X_k} = \{\mathbf{x_k^1}, \mathbf{x_k^2}, ..., \mathbf{x_k^{n_k}}\} \in I\!R^{n_k \times d}, (k = 1, ..., K)$, where $\mathbf{x_k^i} \in I\!R^d$ is a d-dimensional feature vector and n_k is the number of samples in the k-th task. We are going to learn a shared subspace across all tasks, obtained by an orthonormal projection $\mathbf{W} \in I\!R^{d \times s}$, where s is the dimensionality of the subspace. In this learned subspace, the data distribution from all tasks should be similar to each other. Therefore, we can code all tasks together in the shared subspace and achieve better coding quality. The benefits of this strategy are: (i) we can improve each individual coding quality by transferring knowledge across all tasks. (ii) we can discover the relationship among different datasets via coding analysis. Such a purpose can be realized through the following optimization problem:

$$\begin{aligned} \min_{\mathbf{D_k}, \mathbf{C_k}, \mathbf{W}, \mathbf{D}} & \sum_{k=1}^{K} \|\mathbf{X_k} - \mathbf{C_k}\mathbf{D_k}\|_F^2 + \lambda_1 \sum_{k=1}^{K} \|\mathbf{C_k}\|_1 \\ & + \lambda_2 \sum_{k=1}^{K} \|\mathbf{X_k}\mathbf{W} - \mathbf{C_k}\mathbf{D}\|_F^2 \\ s.t. \quad & \begin{cases} \mathbf{W^T W} = \mathbf{I} \\ (\mathbf{D_k})_{\mathbf{j}\cdot} (\mathbf{D_k})_{\mathbf{j}\cdot}^{\mathbf{T}} \leq 1, \quad \forall j = 1, ..., l \\ \mathbf{D_j} \cdot \mathbf{D_j^T} \leq 1, \quad \forall j = 1, ..., l \end{cases} \end{aligned} \tag{1}$$

where $\mathbf{D_k} \in I\!R^{l \times d}$ is an overcomplete dictionary ($l > d$) with l prototypes of the k-th task, $(\mathbf{D_k})_{\mathbf{j}\cdot}$ in the constraints denotes the j-th row of $\mathbf{D_k}$, and $\mathbf{C_k} \in I\!R^{n_k \times l}$ corresponds to the sparse representation coefficients of $\mathbf{X_k}$. In the third term of Eqn.(1), $\mathbf{X_k}$ is projected by $\mathbf{W}$ into the subspace to explore the relationship among different tasks. $\mathbf{D} \in I\!R^{l \times s}$ is the dictionary learned in the datasets shared subspace. $\mathbf{D_j}$. in the constraints denotes the j-th row of $\mathbf{D}$. $\mathbf{I}$ is the identity matrix. $(\cdot)^T$ denotes the transpose operator. λ_1 and λ_2 are the regularization parameters. The first constraint guarantees the learned $\mathbf{W}$ to be orthonormal, and the second and third constraints prevent the learned dictionary to be arbitrarily large. In our objective function, we learn a dictionary $\mathbf{D_k}$ for each task k and one shared dictionary $\mathbf{D}$ among k tasks. Since one task in our model uses samples from the SIN dataset of selected semantic meaningful concepts, the shared learned dictionary $\mathbf{D}$ is the event oriented dictionary.

When $\lambda_2 = 0$, Eqn.(1) reduces to the traditional dictionary learning on separated tasks.

Supervised Multi-task Dictionary Learning

It is well-known that the traditional dictionary learning framework is not directly available for classification and the learned dictionary has merely been used for signal reconstruction (Mairal et al. 2008). To circumvent this problem, researchers have developed several algorithms to learn a classification-oriented dictionary in a supervised learning fashion by exploring the label information. In this subsection, we extend our proposed multi-task dictionary learning of Eqn.(1) to be suitable for event detection.

Assuming that the k-th task has m_k classes, the label information of the k-th task is $\mathbf{Y_k} = \{\mathbf{y_k^1}, \mathbf{y_k^2}, ..., \mathbf{y_k^{n_k}}\} \in I\!R^{n_k \times m_k}$ $(k = 1, ..., K)$, $\mathbf{y_k^i} = [0, ..., 0, 1, 0, ..., 0]$ (the position of non-zero element indicates the class). $\mathbf{\Theta_k} \in I\!R^{l \times m_k}$ is the parameter of the k-th task classifier. Inspired by (Zhang and Li 2010), we consider the following optimization problem:

$$\begin{aligned}\min_{\mathbf{D_k},\mathbf{C_k},\mathbf{\Theta_k},\mathbf{W},\mathbf{D}} & \sum_{k=1}^{K} \|\mathbf{X_k} - \mathbf{C_k}\mathbf{D_k}\|_F^2 + \lambda_1 \sum_{k=1}^{K} \|\mathbf{C_k}\|_1 \\ & + \lambda_2 \sum_{k=1}^{K} \|\mathbf{X_k}\mathbf{W} - \mathbf{C_k}\mathbf{D}\|_F^2 + \lambda_3 \sum_{k=1}^{K} \|\mathbf{Y_k} - \mathbf{C_k}\mathbf{\Theta_k}\|_F^2 \\ s.t. \quad & \begin{cases} \mathbf{W^T W} = \mathbf{I} \\ (\mathbf{D_k})_{\mathbf{j}\cdot}(\mathbf{D_k})_{\mathbf{j}\cdot}^{\mathbf{T}} \le 1, \quad \forall j = 1, ..., l \\ \mathbf{D_j}.\mathbf{D_j^T} \le 1, \quad \forall j = 1, ..., l \end{cases}\end{aligned} \tag{2}$$

Compared with Eqn.(1), we add the last term into Eqn.(2) to enforce the model involving discriminative information for classification. This objective function can simultaneously achieve a desired dictionary with good representation power and support optimal discrimination of the classes for multi-task setting.

To solve the proposed problem of Eqn.(2), we adopt the alternating minimization algorithm to optimize it with respect to $\mathbf{D}$, $\mathbf{D_k}$, $\mathbf{C_k}$, $\mathbf{\Theta_k}$ and $\mathbf{W}$ respectively. We summarize our algorithm for solving Eqn.(2) as Algorithm 1.

After the optimized $\mathbf{\Theta}$ is obtained, the final classification of a test video can be obtained based on its sparse coefficient $\mathbf{c_k^i}$, which delivers the discriminative information. We can simply apply the linear classifier $\mathbf{c_k^i}\mathbf{\Theta_k}$ to obtain the predicted score of the video.

Algorithm 1: Supervised Multi-task Dictionary Learning.

Input:
K tasks Data $(\mathbf{X_1}, ..., \mathbf{X_k})$ and Label $(\mathbf{Y_1}, ..., \mathbf{Y_k})$;
Subspace dimensionality s, Dictionary size l, Regularization parameters λ_1, λ_2, λ_3.

Output:
Optimized $\mathbf{W} \in \mathbb{R}^{d \times s}$, $\mathbf{C_k} \in \mathbb{R}^{n_k \times l}$, $\mathbf{D_k} \in \mathbb{R}^{l \times d}$, $\mathbf{D} \in \mathbb{R}^{l \times s}$, $\mathbf{\Theta_k} \in \mathbb{R}^{l \times m_k}$.

1: Initialize $\mathbf{W}$ using any orthonormal matrix;
2: Initialize $\mathbf{C_k}$ with l_2 normalized columns;
3: **repeat**
Compute $\mathbf{D}$ using Algorithm 2 in (Mairal et al. 2009);
for $k = 1 : K$
Compute $\mathbf{D_k}$ using Algorithm 2 in (Mairal et al. 2009);
Adopting FISTA (Beck and Teboulle 2009) to solve $\mathbf{C_k}$;
$\mathbf{\Theta_k} = (\mathbf{C_k^T C_k})^{-1}\mathbf{C_k^T Y_k}$;
end for
Compute $\mathbf{W}$ by eigen decomposition of $\mathbf{X^T}(\mathbf{I} - \mathbf{C}(\mathbf{C^T C})^{-1}\mathbf{C^T})\mathbf{X}$;
until *Convergence*;

Experiments

Datasets

TRECVID MED10 (P001-P003) and MED11 (E001-E015) datasets are used in our experiments. The datasets consist of 9746 videos from 18 events of interest, with 100-200 examples per event, and the rest of the videos are from the background class.

TRECVID Semantic Indexing Task (SIN) contains annotation for 346 semantic concepts on 400,000 keyframes from web videos. 346 concepts are related to objects, actions, scences, attributes and non-visual concept which are all the basic elements for an event, *e.g.* kitchen, boy, girl, bus. For the sake of better understanding and easy concept selection, we manually divide the 346 visual concepts into 15 groups.

Experiment Setup

There are 3104 videos used for training and 6642 videos used for testing in our experiments. We use three representative features which are SIFT, Color SIFT (CSIFT) and Motion SIFT (MOSIFT) (Chen and Hauptmann 2009). SIFT and CSIFT describe the gradient and color information of images. MOSIFT describes both the optical flow and gradient information of video clips. Finally, 768-dimensional SIFT-BoW, CSIFT-BoW, MOSIFT-BoW features are extracted respectively to represent each video. We set the regularization parameters in the range of $\{0.01, 0.1, 1, 10, 100\}$. The subspace dimensionality s is set by searching the grid from {200, 400, 600}. For the experiments in the paper, we try three different dictionary sizes from {768, 1024, 1280}.

Comparison Method

We compare our proposed event oriented dictionary learning method with the following important baselines:

- *Support Vector Machine (**SVM**)*: SVM has been widely used by several research groups for MED and has shown its robustness (Lan et al. 2013; Oneata et al. 2012), so we use it as one of the comparison algorithms;
- *Single Task Supervised Dictionary Learning (**ST-SDL**)*: Performing supervised dictionary learning on each task separately;
- *Pooling Tasks Supervised Dictionary Learning (**PT-SDL**)*: Performing single task supervised dictionary learning by simply aggregating data from all tasks;
- *Multiple Kernel Transfer Learning (**MKTL**)* (Jie, Tommasi, and Caputo 2011): A method which incorporates prior features into a multiple kernel learning framework;

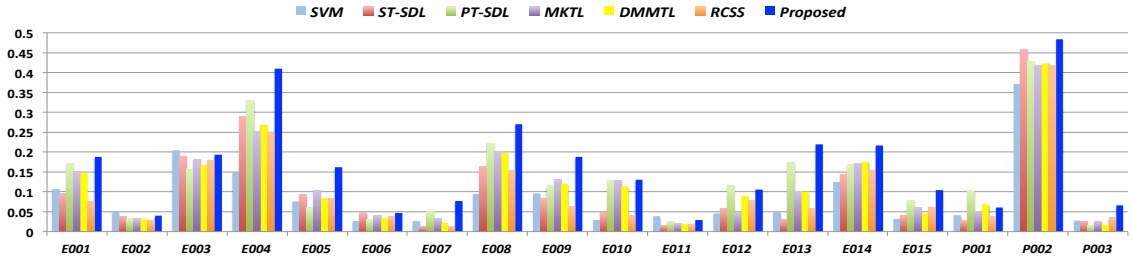

Figure 4: Comparision of different methods of AP performance for each MED event.

Table 1: AP performance for each MED event using Text (T), Visual (V) and Text+Visual (T+V) information for concept selection. The last column shows the number of concepts that are coincided with groundtruth in top 10 concepts.

	Event	T	V	T + V	# in Top 10
P001:	**Assembling shelter**	0.0331	0.0532	**0.0613**	5
P002:	**Batting a run**	0.4432	0.4653	**0.4837**	7
P003:	**Making a cake**	0.0502	0.0514	**0.0658**	9
E001:	**Attempting board trick**	0.1457	0.1675	**0.1883**	6
E002:	**Feeding animal**	0.0355	0.0361	**0.0402**	5
E003:	**Landing fish**	0.1721	0.1801	**0.1938**	5
E004:	**Wedding ceremony**	0.3777	0.3831	**0.4104**	10
E005:	**Working wood working project**	0.1352	0.1542	**0.1625**	9
E006:	**Birthday party**	0.0308	0.0331	**0.0475**	5
E007:	**Changing a vehicle tire**	0.0509	0.0512	**0.0771**	7
E008:	**Flash mob gathering**	0.2433	0.2653	**0.2709**	8
E009:	**Getting a vehicle unstuck**	0.1652	0.1765	**0.1876**	9
E010:	**Grooming an animal**	0.1234	0.1193	**0.1308**	5
E011:	**Making a sandwich**	0.014	0.0213	**0.0285**	4
E012:	**Parade**	0.0761	0.0876	**0.1052**	4
E013:	**Parkour**	0.1769	0.1981	**0.22**	7
E014:	**Repairing an appliance**	0.1742	0.1951	**0.2167**	8
E015:	**Working on a sewing project**	0.071	0.0909	**0.1051**	7

- *Dirty Model Multi-Task Learning (**DMMTL**)* (Jalali et al. 2010): A state-of-the-art multi-task learning method imposing ℓ_1/ℓ_q-norm regularization;
- *Multiple Kernel Learnig Latent Variable Approach* (**MKLLVA**) (Vahdat et al. 2013): A multiple kernel learning latent variable approach for complex video event detection;
- *Random Concept Selection Strategy (**RCSS**)*: Performing our proposed supervised multi-task dictionary learning *without* involving concept selection strategy (leveraging random samples).

Results

To exploit the effectiveness of our proposed concept selection strategy, we compare our selected top 10 concepts with the groundtruth (we use human labeled concepts ranking list as the groundtruth for each MED event). The results are listed in the last column of Table 1, which shows the number of concepts that are coincided with groundtruth in top 10 concepts. Moveover, the AP performance for event detection based on text information, visual information and their combinations are also shown in Table 1. The benefit of using both text and visual information for concept selection can be concluded from Table 1.

Fig.4 shows the AP results for each MED event. Our proposed method achieves the best performance for 13 events out of a total of 18 events. It is also interesting to notice that the larger improvements in Fig.4, such as 'E004: Wedding ceremony', 'E005: Working wood working project' and 'E009: Getting a vehicle unstuck' usually correspond to the higher number of selected concepts that are coincided with groundtruth as shown in Table 1. This gives us the evidence of the effectiveness of proposed automatical concept selection strategy.

Table 2: Comparision of different methods for *average* detection accuracy of SIFT feature.

Method	**MAP**
SVM	0.0883
ST-SDL	0.1037
PT-SDL	0.1336
MKTL (Jie, Tommasi, and Caputo 2011)	0.1191
DMMTL (Jalali et al. 2010)	0.1180
MKLLVA (Vahdat et al. 2013)	0.1132
RCSS	0.1201
Proposed	**0.1664**

Table 2 shows the *average* detection results of the 18 MED events for different comparison methods. We have the following observations: (1) Comparing ST-SDL with SVM, we observe that performing supervised dictionary learning is better than SVM which shows the effectiveness of dictionary learning for MED. (2) Comparing PT-SDL with ST-SDL, leveraging knowledge from the SIN dataset improves the performance for MED. (3) Our concept selection strategy for semantic dictionary learning performs the best for MED among all the comparison methods. (4) Our proposed method outperforms 8% AP compared with SVM. (5) Considering the difficulty of MED dataset and the typically low AP performance of MED, the absolute 8% AP improvement is very significant.

Conclusion

In this paper, we have firstly investigated the possibility of automatically selecting semantic meaningful concepts for complex event detection based on both the MED events-kit text descriptions and the high-level concept feature descriptions. Then we attempt to learn an event oriented dictionary representation based on the selected semantic concepts. To this aim, we leverage training samples of selected concepts from the SIN dataset into a novel jointly supervised multi-task dictionary learning framework. Extensive experimental results on MED dataset show the efficacy of our proposed

semantic concept selection strategy and the event oriented dictionary learning method for complex event detection.

Acknowledgements

This work was partially supported by the MIUR Cluster project Active Ageing at Home and the EC project xLiMe. The work was also supported in part by the U. S. Army Research Office (W911NF-13-1-0277) and by the National Science Foundation under Grant No. IIS-1251187. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of ARO, the National Science Foundation or the U.S. Government.

References

Aharon, M.; Elad, M.; and Bruckstein, A. 2006. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. on Image Processing* 54(11):4311–4322.

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2007. Multi-task feature learning. In *NIPS*.

Beck, A., and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* 2(1):183–202.

Berg, A.; Deng, J.; Satheesh, S.; Su, H.; and Li, F.-F. 2011. Imagenet large scale visual recognition challenge.

Chen, M. Y., and Hauptmann, A. 2009. Mosift: Recognizing human actions in surveillance videos. In *CMU Technical Report, CMU-CS-09-161*.

Elad, M., and Aharon, M. 2006. Image denoising via sparse and redundant representation over learned dictionaries. *IEEE Trans. on Image Processing* 15(12):3736–3745.

Fellbaum, C. 1998. Wordnet: An electronical lexical database. In *The MIT Press, Cambridge, MA*.

Jalali, A.; Ravikumar, P.; Sanghavi, S.; and Ruan, C. 2010. A dirty model for multi-task learning. In *NIPS*.

Jiang, L.; Hauptmann, A. G.; and Xiang, G. 2012. Leveraging high-level and low-level features for multimedia event detection. In *ACM MM*.

Jie, L.; Tommasi, T.; and Caputo, B. 2011. Multiclass transfer learning from unconstrained priors. In *ICCV*.

Lan, Z.-Z.; Jiang, L.; Yu, S.-I.; Rawat, S.; Yang Cai, C. G.; Xu, S.; Shen, H.; Li, X.; Wang, Y.; Sze, W.; Yan, Y.; Ma, Z.; Tong, W.; Yang, Y.; Burger, S.; Metze, F.; Singh, R.; Raj, B.; Stern, R.; Mitamura, T.; Nyberg, E.; and Hauptmann, A. 2013. CMU-Informedia at TRECVID 2013 multimedia event detection. In *Proc. TRECVID 2013 Workshop, Gaithersburg, MD, USA*.

Lee, H.; Battle, A.; Raina, R.; and Ng, A. Y. 2006. Efficient sparse coding algorithms. In *NIPS*.

Lin, D. 1998. An information-theoretic definition of similarity. In *ICML*.

Luo, J.; Papin, C.; and Costello, K. 2009. Towards extracting semantically meaningful key frames from personal video clips: From humans to computers. *IEEE Trans. on circuits and systems for video technology* 19:289–301.

Ma, Z.; Yang, Y.; Xu, Z.; Yan, S.; Sebe, N.; and Hauptmann, A. G. 2013. Complex event detection via multi-source video attributes. In *CVPR*.

Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; and Zisserman, A. 2008. Discriminative learned dictionaries for local image analysis. In *CVPR*.

Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2009. Online dictionary learning for sparse coding. In *ICML*.

Maurer, A.; Pontil, M.; and Paredes, B. R. 2013. Sparse coding for multitask and transfer learning. In *ICML*.

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

Oneata, D.; Douze, M.; Revaud, J.; Jochen, S.; Potapov, D.; Wang, H.; Harchaoui, Z.; Verbeek, J.; Schmid, C.; Aly, R.; Mcguiness, K.; Chen, S.; O'Connor, N.; Chatfield, K.; Parkhi, O.; Arandjelovic, R.; Zisserman, A.; Basura, F.; and Tuytelaars., T. 2012. AXES at TRECVid 2012: KIS, INS, and MED. In *NIST TRECVID Workshop*.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *IJCAI*.

Smeaton, A.; Over, P.; and Kraaij, W. 2006. Evaluation campaigns and trecvid. In *ACM MIR*.

Snoek, C. G. M., and Smeulders, A. W. M. 2010. Visual-concept search solved? *IEEE Computer* 43(6):76–78.

Strube, M., and Ponzetto, S. P. 2006. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*.

Sun, C., and Nevatia, R. 2013. Active: Activity concept transitions in video event classification. In *ICCV*.

Tamrakar, A.; Ali, S.; Yu, Q.; Liu, J.; Javed, O.; Divakaran, A.; Cheng, H.; and Sawhney, H. 2012. Evaluation of low-level features and their combinations for complex event detection in open source videos. In *CVPR*.

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

Yan, Y.; Ricci, E.; Subramanian, R.; Lanz, O.; and Sebe, N. 2013a. No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion. In *ICCV*.

Yan, Y.; Liu, G.; Ricci, E.; and Sebe, N. 2013b. Multi-task linear discriminant analysis for multi-view action recognition. In *ICIP*.

Yang, J.; Yu, K.; Gong, Y.; and Huang, T. 2009. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*.

Zhang, Q., and Li, B. 2010. Discriminative K-SVD for dictionary learning in face recognition. In *CVPR*.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 67:301–320.