Event Oriented Dictionary Learning for Complex Event Detection

Yan Yan, Yi Yang, Deyu Meng, *Member, IEEE*, Gaowen Liu, Wei Tong, Alexander G. Hauptmann, and Nicu Sebe, *Senior Member, IEEE*

***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 a novel strategy 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. Toward this goal, we leverage training images (frames) of selected concepts from the semantic indexing dataset with a pool of 346 concepts, into a novel supervised multitask ℓ_p-norm dictionary learning framework. Extensive experimental results on TRECVID multimedia event detection dataset demonstrate the efficacy of our proposed method.**

***Index Terms*— Complex event detection, concept selection, event oriented dictionary learning, supervised multi-task dictionary learning.**

I. Introduction

COMPLEX event detection in unconstrained videos has received much attention in the research community recently [1]–[3]. 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 from videos, such as 'walking' or 'jumping', 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 the classification outputs of multiple low-level features [1], *i.e.* SIFT, STIP, MOSIFT [4]. Recently, representing videos using high-level features, such as concept detectors [5], appears promising for the complex event detection task. However, the state-of-the-art concept detector based approaches for complex event detection have not considered which concepts should be included in the training concept list. This induces the redundancy of concepts [2], [6] in the concept list for the vocabulary construction. For example, it is highly improbable for some concepts to help detecting a certain event, *e.g.* 'cows' or '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 tends to eliminate such redundancy and potentially boosts the complex event detection performance.

Intuitively, it is highly expected that complex event detection is more accurate and faster when we build a 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 important issues to be considered for accomplishing 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 the 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 introduce the abbreviations used in the paper. SIN stands for the Sematic Indexing dataset [7] containing 346 different categories (concepts)

Manuscript received July 7, 2014; revised October 17, 2014 and February 11, 2015; accepted March 4, 2015. This work was supported in part by the MIUR Cluster Project Active Ageing at Home, in part by the European Commission Project xLiMe, in part by the Australian Research Council Discovery Projects, in part by the U.S. Army Research Office under Grant W911NF-13-1-0277, and in part by the National Science Foundation under Grant IIS-1251187. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gang Hua.

Y. Yan, G. Liu, and N. Sebe are with the Department of Information Engineering and Computer Science, University of Trento, Trento 38123, Italy (e-mail: yan@disi.unitn.it; gaowen.liu@unitn.it; sebe@disi.unitn.it).

Y. Yang is with the Centre for Quantum Computation and Intelligent Systems, University of Technology at Sydney, Sydney, NSW 2007, Australia (e-mail: yee.i.yang@gmail.com).

D. Meng is with the Institute for Information and System Sciences, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: dymeng@mail.xjtu.edu.cn).

W. Tong is with the General Motors Research and Development Center, Warren, MI 48090 USA (e-mail: tongweig@gmail.com).

A. G. Hauptmann is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: alex@cs.cmu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2413294

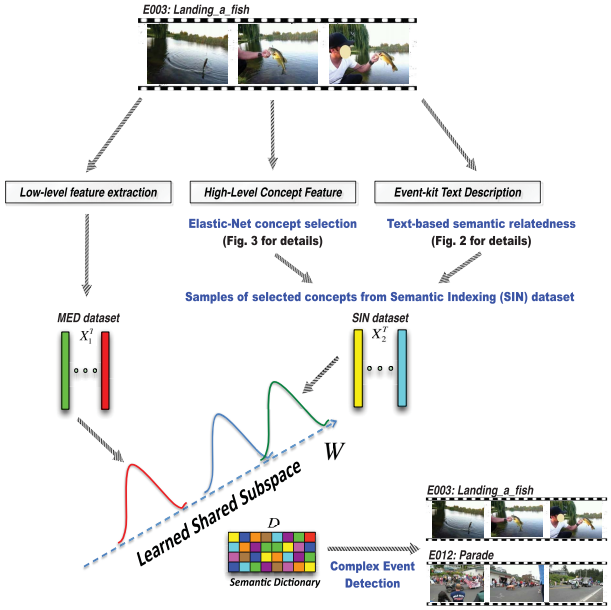


Fig. 1. The event oriented dictionary learning framework.

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 346D feature (each dimension represents a concept).

The overview of our framework is shown in Fig. 1. Firstly, we design a novel method 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 effective complex event detection. Moreover, a novel ℓ_p-norm multi-task dictionary learning is proposed to strengthen the flexibility of the traditional ℓ_1-norm dictionary learning problem.

To summarize, the contributions of this paper are as follows:

- We propose a novel approach for concept selection and present one of the first works making a comprehensive evaluation of automatic concept selection strategies for event detection;
- We propose the event oriented dictionary learning for event detection;
- We construct a supervised multi-task dictionary learning framework which is capable of learning an event oriented dictionary via leveraging information from selected semantic concepts;
- We propose a novel ℓ_p-norm multi-task dictionary learning framework which is more flexible than the traditional ℓ_1-norm dictionary learning problem.

II. Related Work

To highlight our research contributions, we now review the related work on (a) Event Detection, (b) Dictionary Learning and (c) Multi-task Learning.

A. Event Detection

With the success of event detection in structured videos, complex event detection from general *unconstrained* videos, such as those obtained from internet video sharing web sites like YouTube, has received increasing attention in recent years. Unlike traditional action recognition from videos of atomic actions, such as 'walking' or 'jumping', complex event detection aims to detect more complex events such as 'Birthday party', 'Attempting board trick', 'Changing a vehicle tire', *etc*. Tamrakar *et al.* [1] and Lan *et al.* [8] evaluated different low-level appearance as well as spatio-temporal features, appropriately quantized and aggregated into Bag-of-Words (BoW) descriptors for NIST TRECVID Multimedia Event Detection. Jiang *et al.* [9] proposed a method for high-level and low-level feature fusion based on collective classification from three steps which are training a classifier from low-level features, encoding high-level features into graphs, and diffusing the scores on the established graph to obtain the final prediction. Natarajan *et al.* [10] 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 event detection. They combined multiple features using a multi-stage feature fusion strategy with feature level early fusion using multiple kernel learning (MKL) and score level fusion using Bayesian model combination (BayCom) and weighted average fusion using video specific weights. Tang *et al.* [11] tackled the problem of understanding the temporal structure of complex events in highly varying videos obtained from the Internet. A conditional model was trained in a max-margin framework able to automatically discover discriminative and interesting segments of video, while simultaneously achieving competitive accuracies on difficult detection and recognition tasks.

Recently, representing video in terms of multi-model low-level features, *e.g.* SIFT, STIP, Dense Trajectory, Mel-Frequency Cepstral Coefficients (MFCC), Automatic Speech Recognition (ASR), Optical Character Recognition (OCR), combined with early or late fusion schemes is the state-of-the-art [12] for event detection. Despite of their good performance, low-level features are incapable of capturing the inherent semantic information in an event. Comparatively, high-level concept features were shown to be promising for event detection [5]. High-level concept representation approaches become available nowadays due to the availability of large labeled training collections such as ImageNet and TRECVID. However, currently there are still few research works on how to automatically select useful concepts for event detection. Oh *et al.* [13] used Latent SVMs for concept weighting and Brown *et al.* [14] ordered concepts by their discrimination power for each event kit query. Different from [13] and [14], we focus on learning an event oriented dictionary from the perspective of adaptation of the selecting concepts.

B. Dictionary Learning

Dictionary learning (also called Sparse Coding) has been shown 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. Yang *et al.* [15] proposed a spatial pyramid matching approach based on SIFT sparse codes for *image classification*. The method used selective sparse coding instead of the traditional vector quantization to extract salient properties of appearance descriptors of local image patches. Elad and Aharon [16] addressed the *image denoising* problem, where zero-mean white and homogeneous Gaussian additive noise was to be removed from a given image. The approach taken was based on sparse and redundant representations over trained dictionaries. Using the K-SVD algorithm, the authors obtained a dictionary that described the image content effectively. For the *image segmentation* problem, Mairal *et al.* [17] proposed an energy formulation with both sparse reconstruction and class discrimination components, jointly optimized during dictionary learning. The approach improved over the state of the art in image segmentation experiments.

Different optimization algorithms have also been proposed to solve dictionary learning problems. Aharon *et al.* [18] proposed a novel K-SVD algorithm for adapting dictionaries in order to achieve sparse signal representations. K-SVD is an iterative method that alternates between sparse coding of the examples based on the current dictionary and a process of updating the dictionary atoms to better fit the data. The update of the dictionary columns was combined with an update of the sparse representations, thereby accelerating the convergence. Lee *et al.* [19] presented efficient sparse coding algorithms that were based on iteratively solving two convex optimization problems: an ℓ_1-regularized least squares problem and an ℓ_2-constrained least squares problem. To learn a discriminative dictionary for sparse coding, a label consistent K-SVD (LC-KSVD) algorithm was proposed in [20]. In addition to using class labels of training data, the authors also associated label information with each dictionary item (columns of the dictionary matrix) to enforce discriminability in sparse codes during the dictionary learning process. More specifically, a new label consistent constraint was introduced and combined with the reconstruction error and the classification error to form a unified objective function. To effectively handle very large training sets and dynamic training data changing over time, Mairal *et al.* [21] proposed an online optimization algorithm for dictionary learning, based on stochastic approximations, which scaled up to large datasets with millions of training samples.

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.

C. Multi-Task Learning

Multi-task learning [22] methods aim to simultaneously learn classification/regression models for a set of related tasks. This typically leads to better models as compared to a learner that does not account for task relationships. To capture the task relatedness from multiple related tasks is to constrain all models to share a common set of features. This motivates the group sparsity, *i.e.* the $\ell_{2,p}$-norm regularized learning [23]. The joint feature learning using $\ell_{2,p}$-norm regularization performs well in ideal cases. In practical applications, however, simply using the $\ell_{2,p}$-norm regularization may not be effective for dealing with dirty data which may not fall into a single structure. To this end, the dirty model for multi-task learning was proposed in [24]. Another way to capture the task relationship is to constrain the models from different tasks to share a low-dimensional subspace by the trace norm [25]. The assumption that all models share a common low-dimensional subspace is too restrictive in some applications. To this end, an extension that learns incoherent sparse and low-rank patterns simultaneously was proposed in [26].

Many multi-task learning algorithms assume that all learning tasks are related. In practical applications, however, the tasks may exhibit a more sophisticated group structure where the models of tasks from the same group are closer to each other than those from a different group. There have been many works along this line of research [27], [28], known as clustered multi-task learning (CMTL). Moreover, most multi-task learning formulations assume that all tasks are relevant, which is however not the case in many real-world applications. Robust multi-task learning (RMTL) is aimed at identifying irrelevant (outlier) tasks when learning from multiple tasks [29].

However, there is little work on multi-task learning used for dictionary learning problem. The only related theoretical work is that in [30], where only theoretical bounds are provided on evaluating the generalization error of dictionary learning for multi-task learning and transfer learning. Multi-task learning has received considerable attention in the computer vision community and has been successfully applied to many computer vision problems, such as image classification [31], head-pose estimation [32], visual tracking [33], multi-view action recognition [34] and egocentric activity recognition [35]. However, to our knowledge, no previous works have considered the problem of complex event detection.

III. Building an 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. Since the availability of large labeled training collections such as ImageNet and TRECVID, there exists a large pool of concept detectors for event descriptions. However, selecting important concepts is the key issue for concept vocabulary construction. For example, the event 'Landing a fish' is composed of concepts such as 'adult', 'waterscape', 'outdoor' and 'fish'. If concepts related to the event can be accurately selected, redundancy and unrelated information will be suppressed, potentially improving the event recognition performance. In order to select useful concepts for the specific event, we propose a novel concept selection strategy based on the combination of text and visual information provided by the event-kit descriptions, which are (i) Text-based semantic relatedness from linguistic knowledge

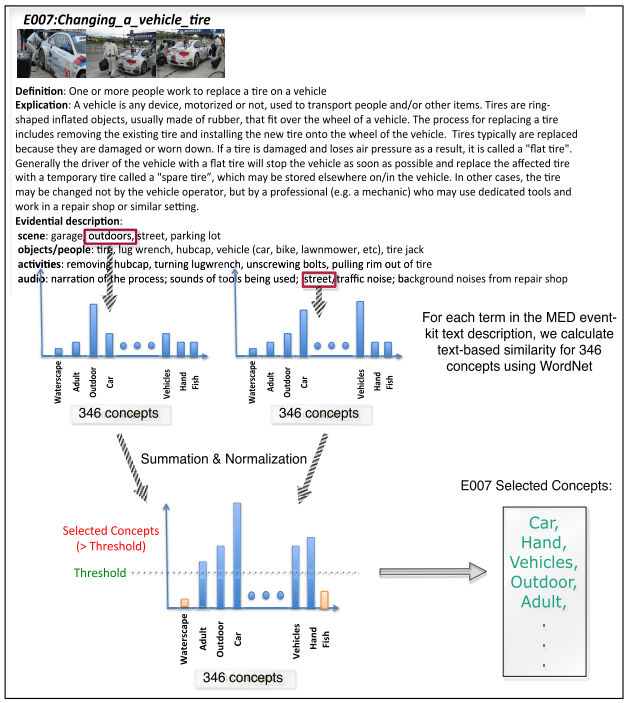


Fig. 2. Linguistic-based concept selection strategy with an example of 'E007: Changing a vehicle tire' in MED event-kit text description and a corresponding example video provided by NIST.

of MED event-kit text description and (ii) Elastic-Net feature selection from visual high-level representation.

A. Linguistic: Text-Based Semantic Relatedness

The most widely used resources in Natural Language Processing (NLP) to calulate the semantic relateness of concepts are WordNet [36] and Wikipedia [37]. There are detailed event-kit text descriptions for each MED event provided by NIST [38]. In this paper, we explore the semantic similarity between each term in the event-kit text description and the SIN 346 visual concept names based on WordNet. Fig. 2 shows an example of event-kit text description for 'Changing a vehicle tire'.

Intuitively, the more information two concepts share in common, the more similar they are, and the information shared by two concepts is indicated by the information content of the concepts that subsume them in the taxonomy. 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 [39]. 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 $w_{1i} \in$ {event-kit text descriptions}, $i = 1, \ldots, N^{event\_kit}$ and $w_{2j} \in$ {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

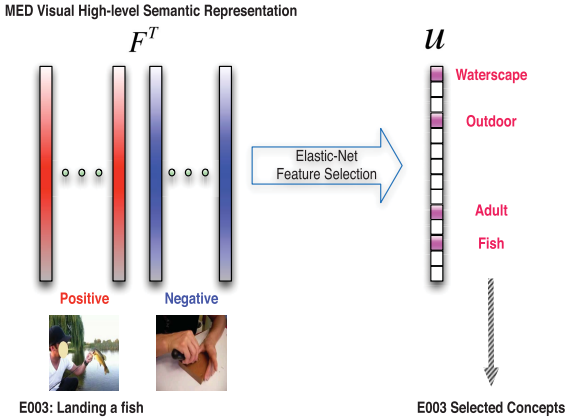


Fig. 3. Visual high-level semantic representation with Elastic-Net concept selection.

$\pi(w) = \log p(w)$, where $p(w)$ is the probability of encountering an instance of w in the union of WordNet and event-kit text descriptions. The probability $p(w) = \text{freq}(w)/N$, which can be estimated from the relative frequency of w and all words N in the union of WordNet and event-kit text descriptions [40]. 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 346D event-level feature vector representation for each event (each dimension corresponds to a SIN visual concept name) using the MED event-kit text description. A threshold is set ($thr = 0.5$ in our experiments) to select useful concepts into our final semantic concept list.

B. Visual High-Level Representation: Elastic-Net Concept Selection

Concept detectors provide a high-level semantic representation for videos with complex contents, which are beneficial for developing powerful retrieval or filtering systems for consumer media [5]. In our case, we firstly use the SIN dataset to train 346 semantic concept models. We adopt the approach [41] 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 346D 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 features.

To select the useful concepts for each specific event, we adopt the Elastic-Net [42] 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 $\mathbf{l} = \{0, 1\}^n \in \boldsymbol{R}^n$ indicates the event labels, $\mathbf{F} \in \boldsymbol{R}^{n \times b}$ is the SIN-MED feature matrix (n is the number of samples

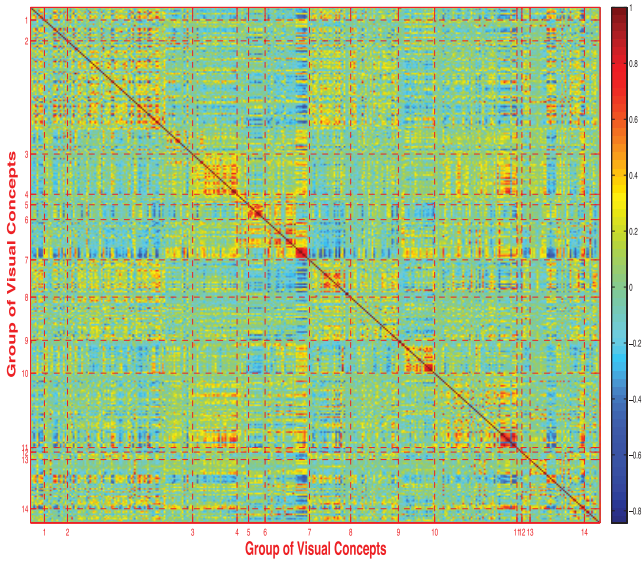


Fig. 4. The correlation demonstration between SIN visual concepts. High correlations between contextually related clusters 'G4:car', 'G7:nature', 'G11:urban-scene' (in red) and negative correlations between contextually unrelated clusters 'G7:nature', 'G8:indoor' (in blue) can be easily observed. (Figure is best viewed in color and under zoom).

and b is the SIN-MED feature dimension) and $\mathbf{u} \in \boldsymbol{R}^b$ 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 [7] (see Fig. 4). 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 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 $\mathbf{u}$.

C. Union of Selected Concepts

To sum up, we form a union of the semantic concepts selected from both text-based semantic relatedness described in section 3.1 and visual high-level semantic representation described in section 3.2 as the final list of selected concepts for each MED event. All the confidence values used are normalized to 0-1 using the Z-score normalization and are used for ranking the concepts. In our paper, we give equal weights to the textual and visual selection methods for the final late fusion of confidence scores. Top 10 ranked concepts are finally selected to adapt semantic information. Since we select the most important concepts from the pool, the top 10 ranked concepts are usually already enough to describe the event (see Fig. 8). To evaluate the effectiveness of our proposed strategy, we compare the selected concepts with the groundtruth (we use human labeled concepts list as the groundtruth for each MED event).

IV. 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.

A. 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 $\mathbf{X_k} = \{\mathbf{x_k^1}, \mathbf{x_k^2}, \ldots, \mathbf{x_k^{n_k}}\} \in \boldsymbol{R}^{n_k \times d}, (k = 1, \ldots, K)$, where $\mathbf{x_k^i} \in \boldsymbol{R}^d$ 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 $\mathbf{W} \in \boldsymbol{R}^{d \times s}$, where s is the dimensionality of the subspace. In this learned subspace, the data distributions 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:

$$\min_{\mathbf{T_k},\mathbf{C_k},\mathbf{W},\mathbf{D}} \sum_{k=1}^{K} \|\mathbf{X_k} - \mathbf{C_k}\mathbf{T_k}\|_F^2 + \lambda_1 \sum_{k=1}^{K} \|\mathbf{C_k}\|_1 + \lambda_2 \sum_{k=1}^{K} \|\mathbf{X_k}\mathbf{W} - \mathbf{C_k}\mathbf{D}\|_F^2$$

$$s.t. \quad \begin{cases} \mathbf{W^T W} = \mathbf{I} \\ (\mathbf{T_k})_{\mathbf{j}\cdot}(\mathbf{T_k})_{\mathbf{j}\cdot}^{\mathbf{T}} \le 1, & \forall j = 1, \ldots, l \\ \mathbf{D_{j\cdot}}\mathbf{D_{j\cdot}^T} \le 1, & \forall j = 1, \ldots, l \end{cases} \quad (1)$$

where $\mathbf{T_k} \in \boldsymbol{R}^{l \times d}$ is an overcomplete dictionary ($l > d$) with l prototypes of the k-th task, $(\mathbf{T_k})_{\mathbf{j}\cdot}$ in the constraints denotes the j-th row of $\mathbf{T_k}$, and $\mathbf{C_k} \in \boldsymbol{R}^{n_k \times l}$ 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. $\mathbf{D} \in \boldsymbol{R}^{l \times s}$ is the dictionary learned in the datasets' shared subspace. $\mathbf{D_{j\cdot}}$ 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 $\mathbf{T_k}$ 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.

B. 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 [17]. 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 $\mathbf{Y_k} = \{\mathbf{y_k^1}, \mathbf{y_k^2}, \ldots, \mathbf{y_k^{n_k}}\} \in \boldsymbol{R}^{n_k \times m_k}$ $(k = 1, \ldots, K)$, $\mathbf{y_k^i} = [0, \ldots, 0, 1, 0, \ldots, 0]$ (the position of non-zero element indicates the class). $\Theta_\mathbf{k} \in \boldsymbol{R}^{l \times m_k}$ is the parameter of the k-th task classifier. Inspired by [43], we consider the following optimization problem:

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

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.

1) Optimization: To solve the proposed objective problem of Eqn.(2), we adopt the alternating minimization algorithm to optimize it with respect to $\mathbf{D}$, $\mathbf{T_k}$, $\mathbf{C_k}$, $\Theta_\mathbf{k}$ and $\mathbf{W}$ respectively in five steps as follows:

Step1 (Fixing $\boldsymbol{T_k}$, $\boldsymbol{C_k}$, $\boldsymbol{W}$, $\boldsymbol{\Theta_k}$, Optimize $\boldsymbol{D}$): If we stack $\mathbf{X} = [\mathbf{X_1^T}, \ldots, \mathbf{X_k^T}]^\mathbf{T}$, $\mathbf{C} = [\mathbf{C_1^T}, \ldots, \mathbf{C_k^T}]^\mathbf{T}$, Eqn.(2) is equivalent to:

$$\begin{aligned}\min_{\mathbf{D}} & \sum_{k=1}^{K} \|\mathbf{X_k}\mathbf{W} - \mathbf{C_k}\mathbf{D}\|_F^2 = \min_{\mathbf{D}} \|\mathbf{XW} - \mathbf{CD}\|_F^2 \\ s.t. & \quad \mathbf{D_j}.\mathbf{D_j^T} \le 1, \quad \forall j = 1, \ldots, l \end{aligned}$$

This is equivalent to the dictionary update stage in traditional dictionary learning algorithm. We adopt the dictionary update strategy of [21, Algorithm 2] to efficiently solve it.

Step2 (Fixing $\boldsymbol{D}$, $\boldsymbol{C_k}$, $\boldsymbol{W}$, $\boldsymbol{\Theta_k}$, Optimize $\boldsymbol{T_k}$): Eqn.(2) is equivalent to:

$$\begin{aligned}\min_{\mathbf{T_k}} & \|\mathbf{X_k} - \mathbf{C_k}\mathbf{T_k}\|_F^2 \\ s.t. & \quad (\mathbf{T_k})_\mathbf{j\cdot}(\mathbf{T_k})_\mathbf{j\cdot}^\mathbf{T} \le 1, \quad \forall j = 1, \ldots, l \end{aligned}$$

This is also equivalent to the dictionary update stage in traditional dictionary learning for k tasks. We adopt the dictionary update strategy of [21, Algorithm 2] to efficiently solve it.

Step3 (Fixing $\boldsymbol{T_k}$, $\boldsymbol{W}$, $\boldsymbol{D}$, $\boldsymbol{\Theta_k}$, Optimize $\boldsymbol{C_k}$): Eqn.(2) is equivalent to:

$$\begin{aligned}\min_{\mathbf{C_k}} & \sum_{k=1}^{K} \|\mathbf{X_k} - \mathbf{C_k}\mathbf{T_k}\|_F^2 + \lambda_1 \sum_{k=1}^{K} \|\mathbf{C_k}\|_1 \\ & + \lambda_2 \sum_{k=1}^{K} \|\mathbf{X_k}\mathbf{W} - \mathbf{C_k}\mathbf{D}\|_F^2 + \lambda_3 \sum_{k=1}^{K} \|\mathbf{Y_k} - \mathbf{C_k}\Theta_\mathbf{k}\|_F^2 \end{aligned}$$

This formulation can be decoupled into $(n_1 + n_2 + \ldots + n_k)$ distinct problems:

$$\begin{aligned}\min_{\mathbf{c_k^i}} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (& \left\|\mathbf{x_k^i} - \mathbf{c_k^i}\mathbf{T_k}\right\|_2^2 + \lambda_1 \left\|\mathbf{c_k^i}\right\|_1 \\ & + \lambda_2 \left\|\mathbf{x_k^i}\mathbf{W} - \mathbf{c_k^i}\mathbf{D}\right\|_2^2 + \lambda_3 \left\|\mathbf{y_k^i} - \mathbf{c_k^i}\Theta_\mathbf{k}\right\|_2^2) \end{aligned}$$

We adopt the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [44] to solve the problem. FISTA solves the optimization problems in the form of $\min_{\boldsymbol{\mu}} f(\boldsymbol{\mu}) + r(\boldsymbol{\mu})$, where $f(\boldsymbol{\mu})$ is convex and smooth, and $r(\boldsymbol{\mu})$ is convex but non-smooth. We adopt FISTA since it is a popular tool for solving many convex smooth/non-smooth problems and its effectiveness has been verified in many applications. In our setting, we denote the smooth term part as $f(\boldsymbol{c_k^i}) = \left\|\mathbf{x_k^i} - \mathbf{c_k^i}\mathbf{T_k}\right\|_2^2 + \lambda_2 \left\|\mathbf{x_k^i}\mathbf{W} - \mathbf{c_k^i}\mathbf{D}\right\|_2^2 + \lambda_3 \left\|\mathbf{y_k^i} - \mathbf{c_k^i}\Theta_\mathbf{k}\right\|_2^2$ and the non-smooth term part as $g(\boldsymbol{c_k^i}) = \lambda_1 \left\|\mathbf{c_k^i}\right\|_1$.

Step4 (Fixing $\boldsymbol{D}$, $\boldsymbol{C_k}$, $\boldsymbol{W}$, $\boldsymbol{T_k}$, Optimize $\boldsymbol{\Theta_k}$): Eqn.(2) is equivalent to:

$$\min_{\Theta_\mathbf{k}} \|\mathbf{Y_k} - \mathbf{C_k}\Theta_\mathbf{k}\|_F^2$$

Setting $\frac{\partial}{\partial \Theta_\mathbf{k}} = 0$, we obtain $\Theta_\mathbf{k} = (\mathbf{C_k^T C_k})^{-1}\mathbf{C_k^T Y_k}$.

Step5 (Fixing $\boldsymbol{T_k}$, $\boldsymbol{C_k}$, $\boldsymbol{D}$, $\boldsymbol{\Theta_k}$, Optimize $\boldsymbol{W}$): If we stack $\mathbf{X} = [\mathbf{X_1^T}, \ldots, \mathbf{X_k^T}]^\mathbf{T}$, $\mathbf{C} = [\mathbf{C_1^T}, \ldots, \mathbf{C_k^T}]^\mathbf{T}$, Eqn.(2) is equivalent to:

$$\begin{aligned}\min_{\mathbf{W}} & \sum_{k=1}^{K} \|\mathbf{X_k}\mathbf{W} - \mathbf{C_k}\mathbf{D}\|_F^2 = \min_{\mathbf{W}} \|\mathbf{XW} - \mathbf{CD}\|_F^2 \\ s.t. & \quad \mathbf{W^T W} = \mathbf{I} \end{aligned}$$

Substituting $\mathbf{D} = (\mathbf{C^T C})^{-1}\mathbf{C^T XW}$ back into the above function, we achieve

$$\begin{aligned}\min_{\mathbf{W}} & \left\|(\mathbf{I} - \mathbf{C}(\mathbf{C^T C})^{-1}\mathbf{C^T})\mathbf{XW}\right\|_F^2 \\ & = \min_{\mathbf{W}} tr(\mathbf{W^T X^T}(\mathbf{I} - \mathbf{C}(\mathbf{C^T C})^{-1}\mathbf{C^T})\mathbf{XW}) \\ s.t. & \quad \mathbf{W^T W} = \mathbf{I} \end{aligned}$$

The optimal $\mathbf{W}$ is composed of eigenvectors of the matrix $\mathbf{X^T}(\mathbf{I} - \mathbf{C}(\mathbf{C^T C})^{-1}\mathbf{C^T})\mathbf{X}$ corresponding to the s smallest eigenvalues.

We summarize our algorithm for solving Eqn.(2) as Algorithm 1.

Algorithm 1 Supervised Multi-Task Dictionary Learning

Input:
K tasks: Data $(\mathbf{X_1}, ..., \mathbf{X_k})$ and Labels $(\mathbf{Y_1}, ..., \mathbf{Y_k})$;
Subspace dimensionality s, Dictionary size l, Regularization parameters λ_1, λ_2, λ_3.
Output:
Optimized $\mathbf{W} \in \mathbb{R}^{d\times s}$, $\mathbf{C_k} \in \mathbb{R}^{n_k\times l}$, $\mathbf{T_k} \in \mathbb{R}^{l\times d}$, $\mathbf{D} \in \mathbb{R}^{l\times s}$, $\mathbf{\Theta_k} \in \mathbb{R}^{l\times m_k}$.

Initialize $\mathbf{W}$ using any orthonormal matrix;
Initialize $\mathbf{C_k}$ with l_2 normalized columns;
repeat
  Compute $\mathbf{D}$ using Algorithm 2 in [21];
  for $k = 1 : K$
    Compute $\mathbf{T_k}$ using Algorithm 2 in [21];
    Adopting FISTA [44] 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 $\mathbf{X^T(I - C(C^TC)^{-1}C^T)X}$;
until *Convergence*;

Algorithm 2 Supervised Multi-Task ℓ_p-Norm Dictionary Learning

Input:
K tasks: Data $(\mathbf{X_1}, ..., \mathbf{X_k})$ and Labels $(\mathbf{Y_1}, ..., \mathbf{Y_k})$;
Subspace dimensionality s, Dictionary size l, ℓ_p-norm parameter p, Regularization parameters λ_1, λ_2, λ_3.
Output:
Optimized $\mathbf{W} \in \mathbb{R}^{d\times s}$, $\mathbf{C_k} \in \mathbb{R}^{n_k\times l}$, $\mathbf{T_k} \in \mathbb{R}^{l\times d}$, $\mathbf{D} \in \mathbb{R}^{l\times s}$, $\mathbf{\Theta_k} \in \mathbb{R}^{l\times m_k}$.

Initialize $\mathbf{W}$ using any orthonormal matrix;
Initialize $\mathbf{C_k}$ with l_2 normalized columns;
repeat
  Compute $\mathbf{D}$ using Algorithm 2 in [21];
  for $k = 1 : K$
    Compute $\mathbf{T_k}$ using Algorithm 2 in [21];
    Adopting GISA [48] 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 $\mathbf{X^T(I - C(C^TC)^{-1}C^T)X}$;
until *Convergence*;

C. Supervised Multi-Task ℓ_p-Norm Dictionary Learning

In the literature it has been shown that using a non-convex ℓ_p-norm minimization ($0 \le p < 1$) can often yield better results than the convex ℓ_1-norm minimization. Inspired by this, we extend our supervised multi-task dictionary learning model to a supervised multi-task ℓ_p-norm dictionary learning model.

Assuming that the k-th task has m_k classes, the label information of the k-th task is $\mathbf{Y_k} = \{\mathbf{y_k^1}, \mathbf{y_k^2}, \ldots, \mathbf{y_k^{n_k}}\} \in \boldsymbol{R}^{n_k\times m_k}$, $(k = 1, \ldots, K)$, $\mathbf{y_k^i} = [0, \ldots, 0, 1, 0, \ldots, 0]$ (the position of non-zero element indicates the class). $\Theta_\mathbf{k} \in \boldsymbol{R}^{l\times m_k}$ is the parameter of the k-th task classifier. We formulate our supervised multi-task ℓ_p-norm dictionary learning problem as follows:

$$\begin{aligned}\min_{\mathbf{T_k},\mathbf{C_k},\Theta_\mathbf{k},\mathbf{W},\mathbf{D}} & \sum_{k=1}^{K}\|\mathbf{X_k} - \mathbf{C_k}\mathbf{T_k}\|_F^2 + \lambda_1\sum_{k=1}^{K}\|\mathbf{C_k}\|_p^p \\ & +\lambda_2\sum_{k=1}^{K}\|\mathbf{X_k}\mathbf{W} - \mathbf{C_k}\mathbf{D}\|_F^2 + \lambda_3\sum_{k=1}^{K}\|\mathbf{Y_k} - \mathbf{C_k}\Theta_\mathbf{k}\|_F^2 \\ s.t. \quad & \begin{cases}\mathbf{W^TW} = \mathbf{I} \\ (\mathbf{T_k})_\mathbf{j\cdot}(\mathbf{T_k})_\mathbf{j\cdot}^\mathbf{T} \le 1, & \forall j = 1, \ldots, l \\ \mathbf{D_{\cdot j}}\mathbf{D_{\cdot j}^T} \le 1, & \forall j = 1, \ldots, l\end{cases}\end{aligned} \tag{3}$$

Compared with Eqn.2, we replace the traditional sparse coding ℓ_1-norm term $\|\mathbf{C_k}\|_1$ with the more flexible ℓ_p-norm term $\|\mathbf{C_k}\|_p^p$. Since we can adjust the value of p ($0 \le p < 1$) in our framework, our algorithm is more flexible to control the sparseness of the feature representation, thus usually resulting in better performance than the traditional ℓ_1-norm sparse coding.

To solve the proposed problem of Eqn.3, we adopt the alternating minimization algorithm to optimize it with respect to $\mathbf{D}, \mathbf{T_k}, \mathbf{C_k}, \Theta_\mathbf{k}, \mathbf{W}$ respectively. The updated rules for $\mathbf{D}, \mathbf{T_k}, \Theta_\mathbf{k}, \mathbf{W}$ are the same as Eqn.2, the only difference is in the updated rule of $\mathbf{C_k}$. Various algorithms have been proposed for ℓ_p-norm non-convex sparse coding [45]–[47]. In this paper, we adopt the Generalized Iterated Shrinkage Algorithm (GISA) [48] to solve the proposed problem. We summarize our algorithm for solving Eqn.3 as Algorithm 2.

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

TABLE I
18 EVENTS OF MED10 AND MED11

	Event Name	Train-set Positive #	Train-set Negative #	Test-set Positive #	Test-set Negative #
P001:	**Assembling shelter**	51	3053	45	6597
P002:	**Batting a run**	54	3050	52	6590
P003:	**Making a cake**	59	3045	47	6595
E001:	**Attempting board trick**	161	2943	114	6528
E002:	**Feeding animal**	162	2942	114	6528
E003:	**Landing fish**	119	1984	85	6557
E004:	**Wedding ceremony**	125	2979	87	6555
E005:	**Working wood working project**	141	2963	99	6543
E006:	**Birthday party**	87	3017	86	6556
E007:	**Changing a vehicle tire**	56	3048	55	6587
E008:	**Flash mob gathering**	87	3017	86	6556
E009:	**Getting a vehicle unstuck**	64	3040	66	6576
E010:	**Grooming an animal**	69	3035	69	6573
E011:	**Making a sandwich**	62	3042	63	6579
E012:	**Parade**	68	3036	69	6573
E013:	**Parkour**	56	3048	55	6587
E014:	**Repairing an appliance**	62	3042	61	6581
E015:	**Working on a sewing project**	60	3044	60	6582

V. EXPERIMENTS

In this section, we conduct extensive experiments to test our proposed method using large-scale real world datasets.

A. 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. The details are listed in the Table 1.

TRECVID Semantic Indexing Task (SIN) [7] contains annotation for 346 semantic concepts on 400,000 keyframes from web videos. 346 concepts are related to objects, actions,

TABLE II
15 GROUPS OF SIN 346 VISUAL CONCEPTS (THE NUMBER OF CONCEPTS FOR EACH GROUP ARE IN PARENTHESIS)

G1:	**Body_Parts** (8)	G2:	**Person** (14)	G3:	**Military** (76)
G4:	**Car** (27)	G5:	**Boat** (7)	G6:	**Aircraft** (10)
G7:	**Nature** (27)	G8:	**Indoor** (25)	G9:	**News** (29)
G10:	**Animal** (22)	G11:	**Urban_Scenes** (50)	G12:	**Natural_Disaster** (3)
G13:	**Election** (5)	G14:	**Sport_Activity** (33)	G15:	**Moods** (10)

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 which are listed in Table 2.

B. Evaluation Metrics

1) Average Precision (AP): is a measure combining recall and precision for ranked retrieval results. The average precision is the mean of the precision scores after each relevant sample is retrieved. The *higher* number indicates the better performance.

2) PMiss@TER = 12.5: is an official evaluation metric for event detection as defined by NIST [38]. It is defined as the point at which the ratio between the probability of Missed Detection and probability of False Alarm is 12.5:1. The *lower* number indicates the better performance.

3) Normalized Detection Cost (NDC): is an official evaluation metric for event detection as defined by NIST [38]. It is a weighted linear combination of the system's Missed Detection and False Alarm probabilities. NDC measures the performance of a detection system in the context of an application profile using error rate estimates calculated on a test set. The *lower* number indicates the better performance.

C. Experiment Settings and Comparison Methods

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) [4]. SIFT and CSIFT describe the gradient and color information of images. MOSIFT describes both the optical flow and gradient information of video clips. Finally, 768D 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 $\{768, 1024, 1280\}$. To evaluate the multi-task ℓ_p-norm dictionary learning algorithm, the parameter p is tuned in the range of $\{0.2, 0.4, 0.6, 0.8, 1\}$. 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 [8], [49], [50], so we use it as one of the comparison algorithms (RBF, Histogram Intersection and χ^2 kernels are used respectively);

Fig. 5. Example results of semantic concept selection proposed in section 3 on event (top left) *Attempting board trick*, (top right) *Feeding animal*, (bottom left) *Flash mob gathering*, (bottom right) *Making a sandwich*. The font size in the figure reflects the ranking value for concepts (the larger the font the higher the value).

TABLE III
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 IN THE TOP 10 THAT COINCIDE WITH THE GROUNDTRUTH

	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

- *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**) [51]:* A method incorporating prior features into a multiple kernel learning framework. We use the code provided by the author[1];
- *Dirty Model Multi-Task Learning (**DMMTL**) [24]:* A state-of-the-art multi-task learning method imposing ℓ_1/ℓ_q-norm regularization. We use the code provided by MALSAR toolbox[2];
- *Multiple Kernel Learnig Latent Variable Approach (**MKLLVA**) [52]:* 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

[1]http://homes.esat.kuleuven.be/~ttommasi/source_code_ICCV11.html
[2]http://www.public.asu.edu/~jye02/Software/MALSAR/

TABLE IV
COMPARISON OF *Average* DETECTION ACCURACY OF DIFFERENT METHODS FOR THE SIFT FEATURE. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Evaluation Metric	SVM_{RBF}	SVM_{HI}	SVM_{Chi2}	ST-SDL	PT-SDL	MKTL [51]	DMMTL [24]	MKLLVA [52]	RCSS	SCSS	Proposed
AP	0.0731	0.0856	0.0883	0.1037	0.1336	0.1191	0.1180	0.1132	0.1201	0.1346	**0.1664**
PMiss@TER=12.5	0.6612	0.6501	0.6535	0.6447	0.6127	0.6364	0.6133	0.6106	0.6221	0.6199	**0.5927**
MinNDC	0.9541	0.9378	0.9401	0.9154	0.8644	0.8843	0.8674	0.8731	0.8612	0.8752	**0.8404**

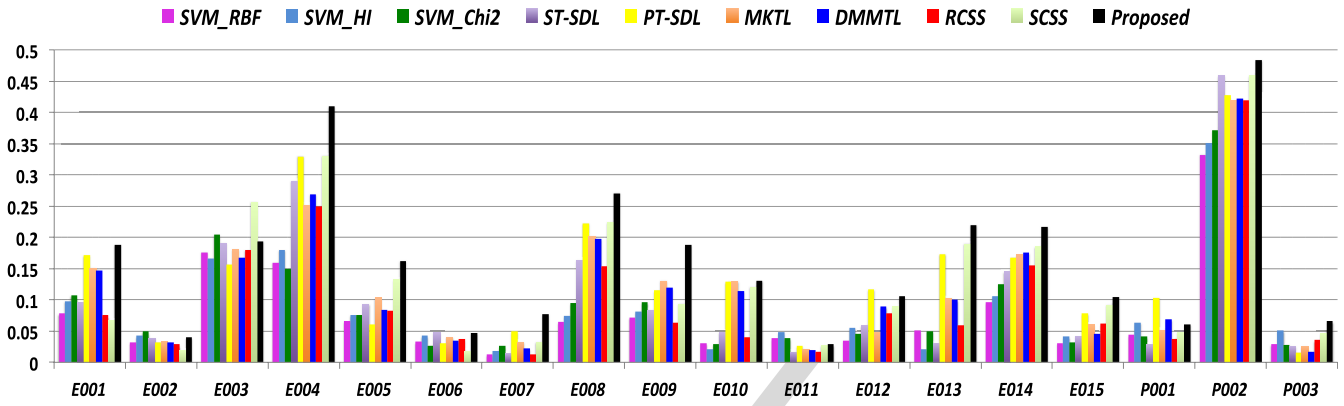


Fig. 6. Comparison of AP performance of different methods for each MED event. (Figure is best viewed in color and under zoom).

without involving concept selection strategy (leveraging random samples).

- *Simple Concept Selection Strategy (**SCSS**):* Each concept is independently used as a predictor for each event yielding a certain average precision performance for each event. The concepts are ranked according to the AP measurement for each event, and the top 10 concepts retained.

D. Experimental Results

We firstly calculate the covariance matrix of the 346 SIN concepts, shown in Fig. 4, where the visual concepts are grouped and sorted the same as in Table 2. As shown in Fig. 4, the convariance matrix shows high within-cluster correlations along the diagonal direction and also relatively high correlations between contextually related clusters (in red color), such as 'G4:car', 'G7:nature' and 'G11:urban-scene'. One can also easily observe negative correlations between contextually unrelated clusters (in blue color), such as 'G7:nature' and 'G8:indoor'. This gives us the intuition that visual concepts co-occurrence exists. Therefore, removing redundant concepts and selecting related concepts for each event is expected to be helpful for the event detection task.

Fig. 5 shows the results of our concept selection strategy for the event 'Attempting board trick', 'Feeding animal', 'Flash mob gathering' and 'Making a sandwich'. From Fig. 5, we can observe that the concepts selected are reasonably consistent with human selections.

To better exploit the effectiveness of our proposed concept selection strategy, we compare our selected top 10 concepts with the groundtruth (we use the ranking list of human labeled concepts as the groundtruth for each MED event). The results are listed in the last column of Table 3, showing the number of concepts in the top 10 that coincide with the groundtruth. The AP performance for event detection based on text information, visual information and their combinations are also shown in Table 3. The benefit of using both text and visual information for concept selection can be concluded from Table 3.

Table 4 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 by 8%, 6%, 10% with respect to AP, PMiss@TER=12.5 and MinNDC respectively 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.

Fig. 6 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. 6, 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 coincide with the groundtruth. This gives us the evidence of the effectiveness of our proposed automatic concept selection strategy.

Fig. 7 illustrates the MAP performance for different methods based on SIFT, CSIFT and MOSIFT features. It can be easily observed that our proposed surpervised multi-task dictionary learning with our concept selection strategy outperforms SVM by more than 8%.

Moreover, we evaluate our proposed method with respect to different numbers of selected concepts, dictionary sizes and

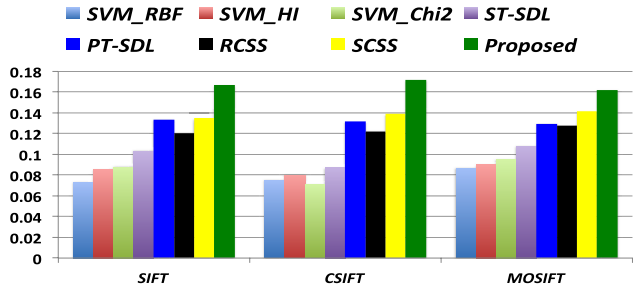


Fig. 7. Comparison of MAP performance of different methods for different types of features.

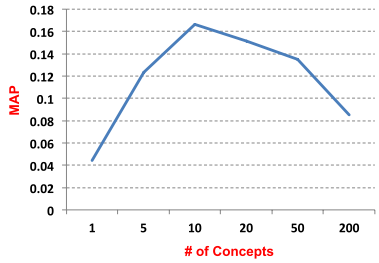


Fig. 8. MAP performance variation with respect to the number of selected concepts.

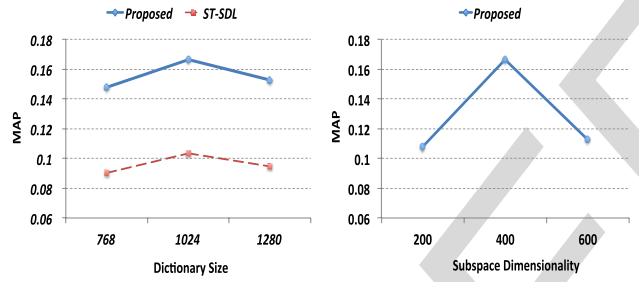


Fig. 9. MAP performance variation with respect to (left) different dictionary size; (right) different subspace dimensionality.

different subspace dimensionality settings based on the SIFT feature. Fig. 8 shows that the proposed method achieves the best MAP results when the numbers of selected concepts is 10. Fig. 9(left) shows that the proposed method achieves the best MAP results when the dictionary size is 1024. Too large or too small dictionary size tends to hamper the performance. Fig. 9(right) shows that the best MAP result is achieved when the subspace dimensionality is 400 (dictionary size = 1024). Large or small subspace dimensionality also degrades the performance.

Finally, we also study the parameter sensitivity of the proposed method in Fig. 10. Here, we fix $\lambda_3 = 1$ (discriminative information contribution fixed) and $p = 0.6$ and analyze the regularization parameters λ_1 and λ_2. As shown in Fig. 10(left), we observe that the proposed method is more sensitive to λ_2 compared with λ_1, which demonstrates the importance of the subspace for multi-task dictionary learning. Moreover, to understand the influence of parameter p for our proposed supervised ℓ_p-norm dictionary learning algorithm, we also perform an experiment on the parameter sensitivity. Fig. 10(right) demonstrates that the best performance for the

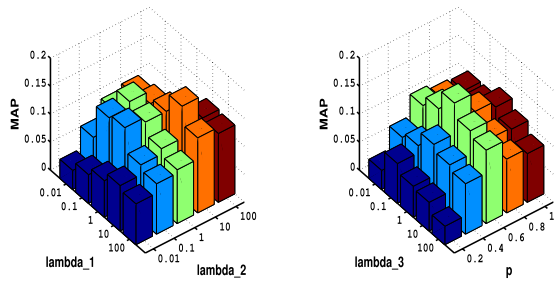


Fig. 10. Sensitivity study of parameters on (left) λ_1 and λ_2 with fixed p and λ_3. (right) p and λ_3 with fixed λ_1 and λ_2.

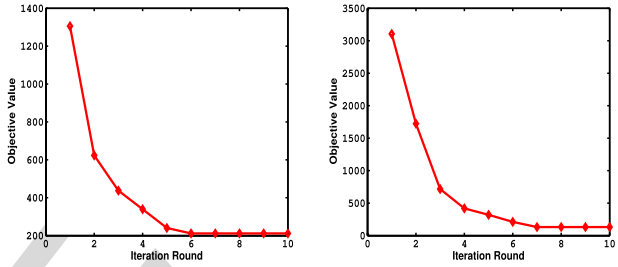


Fig. 11. Convergence rate study on (left) supervised multi-task dictionary learning and (right) supervised multi-task ℓ_p-norm dictionary learning.

supervised ℓ_p-norm dictionary learning algorithm is achieved when $p = 0.6$. More than 2% MAP can be achieved if we adopt the ℓ_p-norm model compared with the fixed ℓ_1-norm model. This shows the suboptimality of the traditional ℓ_1-norm sparse coding compared with the flexible ℓ_p-norm sparse coding.

The proposed iterative approaches monotonically decrease the objective function values in Eqn.(2) and Eqn.(3) until convergence. Fig. 11 shows the convergence rate curves of our algorithms. It can be observed that the objective function values converge quickly and our approaches usually converge after 6 iterations for supervised multi-task dictionary learning and 7 iterations for supervised multi-task ℓ_p-norm dictionary learning at most (precision $= 10^{-6}$).

Regarding the computational cost of our proposed algorithm for supervised multi-task ℓ_p-norm dictionary learning, we train our model for TRECVID MED dataset in 5 hours with cross-validation on a workstation with Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz × 17 processors. Our proposed event oriented dicionary learning approach can be easily parallelled on multi-core computers due to its 'event oriented'. This means that our algorithms would be scalable for large-scale problems.

VI. Conclusion and Future Work

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 the TRECVID MED dataset showed that our proposed method outperformed several important baseline methods. Our proposed method outperformed SVM by up to 8% MAP which showed the effectiveness of dictionary learning for TRECVID MED. More than 6% and 3% MAP was achieved respectively compared with ST-SDL and PT-SDL, which showed the advantage of multi-task setting in our proposed framework. To show the benefit of concept selection strategy, we compared RCSS to our method and showed that achieves 4% less MAP.

For some sparse coding problems, non-convex ℓ_p-norm minimization ($0 \leq p < 1$) can often obtain better results than the convex ℓ_1-norm minimization. Inspired by this, we extended our supervised multi-task dictionary learning model to a supervised multi-task ℓ_p-norm dictionary learning model. We evaluated the influence of the ℓ_p-norm parameter p in our proposed problem and found that more than 2% MAP can be achieved if we adopted the more flexible ℓ_p-norm model compared with the fixed ℓ_1-norm model.

Overall, the proposed multi-task dictionary learning solutions are novel in the context of complex event detection, which is a relevant and important research problem in applications such as image and video understanding and surveillance. Future research involves (i) integration of knowledge from multiple sources (video, audio, text) and incorporation of kernel learning in our framework, and (ii) the use of deep structures instead of a shallow single-layer model in the proposed problem since deep learning has achieved the supreme success in many different fields of image processing and computer vision.

ACKNOWLEDGEMENT

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

[1] A. Tamrakar *et al.*, "Evaluation of low-level features and their combinations for complex event detection in open source videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3681–3688.

[2] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. G. Hauptmann, "Complex event detection via multi-source video attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2627–2633.

[3] Y. Yan, Y. Yang, H. Shen, D. Meng, G. Liu, and N. S. A. Hauptmann, "Complex event detection via event oriented dictionary learning," in *Proc. AAAI Conf. Artif. Intell.*, 2015.

[4] M.-Y. Chen and A. Hauptmann, "MoSIFT: Recognizing human actions in surveillance videos," School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-09-161, 2009.

[5] C. G. M. Snoek and A. W. M. Smeulders, "Visual-concept search solved?" *Computer*, vol. 43, no. 6, pp. 76–78, Jun. 2010.

[6] C. Sun and R. Nevatia, "ACTIVE: Activity concept transitions in video event classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 913–920.

[7] SIN. (2013). *NIST TRECVID Semantic Indexing*. [Online]. Available: http://www-nlpir.nist.gov/projects/tv2013/tv2013.html#sin

[8] Z.-Z. Lan *et al.*, "Informedia E-lamp @ TRECVID 2013 multimedia event detection and recounting (MED and MER)," in *Proc. NIST TRECVID Workshop*, 2013.

[9] L. Jiang, A. G. Hauptmann, and G. Xiang, "Leveraging high-level and low-level features for multimedia event detection," in *Proc. ACM Int. Conf. Multimedia*, 2012, pp. 449–458.

[10] P. Natarajan *et al.*, "Multimodal feature fusion for robust event detection in web videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1298–1305.

[11] K. Tang, D. Koller, and L. Fei-Fei, "Learning latent temporal structure for complex event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1250–1257.

[12] Z.-Z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann, "Double fusion for multimedia event detection," in *Proc. Int. Conf. Multimedia Modeling*, 2012, pp. 173–185.

[13] S. Oh *et al.*, "Multimedia event detection with multimodal feature fusion and temporal concept localization," *Mach. Vis. Appl.*, vol. 25, no. 1, pp. 49–69, Jan. 2014.

[14] L. Brown *et al.*, "IBM Research and Columbia University TRECVID-2013 multimedia event detection (MED), multimedia event recounting (MER), surveillance event detection (SED), and semantic indexing (SIN) systems," in *Proc. NIST TRECVID Workshop*, 2013.

[15] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1794–1801.

[16] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[17] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[18] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[19] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 2006, pp. 801–808.

[20] Z. Jiang, Z. Lin, and L. S. Davis, "Learning a discriminative dictionary for sparse coding via label consistent K-SVD," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1697–1704.

[21] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 689–696.

[22] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 109–117.

[23] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1–8.

[24] A. Jalali, P. K. Ravikumar, S. Sanghavi, and C. Ruan, "A dirty model for multi-task learning," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 2010.

[25] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 457–464.

[26] J. Chen, J. Liu, and J. Ye, "Learning incoherent sparse and low-rank patterns from multiple tasks," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, p. 22.

[27] L. Jacob, F. R. Bach, and J.-P. Vert, "Clustered multi-task learning: A convex formulation," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 2008.

[28] J. Zhou, J. Chen, and J. Ye, "Clustered multi-task learning via alternating structure optimization," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 2011, pp. 702–710.

[29] J. Chen, J. Zhou, and J. Ye, "Integrating low-rank and group-sparse structures for robust multi-task learning," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 42–50.

[30] A. Maurer, M. Pontil, and B. Romera-Paredes, "Sparse coding for multitask and transfer learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 343–351.

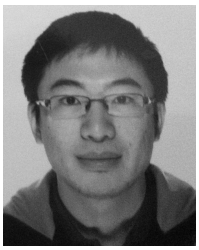
[31] X.-T. Yuan and S. Yan, "Visual classification with multi-task joint sparse representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3493–3500.

[32] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, and N. Sebe, "No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1177–1184.

[33] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2042–2049.

[34] Y. Yan, E. Ricci, R. Subramanian, G. Liu, and N. Sebe, "Multitask linear discriminant analysis for view invariant action recognition," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5599–5611, Dec. 2014.

[35] Y. Yan, E. Ricci, G. Liu, and N. Sebe, "Recognizing daily activities from first-person videos with multi-task clustering," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 1–16.
[36] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. Cambridge, MA, USA: MIT Press, 1998.
[37] M. Strube and S. P. Ponzetto, "WikiRelate! Computing semantic relatedness using Wikipedia," in *Proc. AAAI Conf. Artif. Intell.*, 2006, pp. 1–6.
[38] NIST. (2013). *NIST TRECVID Multimedia Event Detection*. [Online]. Available: http://www.nist.gov/itl/iad/mig/med13.cfm
[39] D. Lin, "An information-theoretic definition of similarity," in *Proc. 5th Int. Conf. Mach. Learn.*, 1998, pp. 296–304.
[40] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. 14th Int. Joint Conf. Artif. Intell.*, 1995, pp. 448–453.
[41] J. Luo, C. Papin, and K. Costello, "Towards extracting semantically meaningful key frames from personal video clips: From humans to computers," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 289–301, Feb. 2009.
[42] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., Ser. B*, vol. 67, no. 2, pp. 301–320, Apr. 2005.
[43] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2691–2698.
[44] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.
[45] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.
[46] Y. She, "Thresholding-based iterative selection procedures for model selection and shrinkage," *Electron. J. Statist.*, vol. 3, pp. 384–415, Nov. 2009.
[47] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-Laplacian priors," in *Proc. Conf. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1033–1041.
[48] W. Zuo, D. Meng, L. Zhang, X. Feng, and D. Zhang, "A generalized iterated shrinkage algorithm for non-convex sparse coding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 217–224.
[49] L. Cao *et al.*, "IBM Research and Columbia University TRECVID-2011 multimedia event detection (MED) system," in *Proc. NIST TRECVID Workshop*, Dec. 2011.
[50] D. Oneata *et al.*, "AXES at TRECVid 2012: KIS, INS, and MED," in *Proc. NIST TRECVID Workshop*, 2012.
[51] L. Jie, T. Tommasi, and B. Caputo, "Multiclass transfer learning from unconstrained priors," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1863–1870.
[52] A. Vahdat, K. Cannons, G. Mori, S. Oh, and I. Kim, "Compositional models for video event detection: A multiple kernel learning latent variable approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1185–1192.



Yan Yan received the Ph.D. degree from the University of Trento, in 2014. He is currently a Post-Doctoral Researcher with the MHUG Group, University of Trento. His research interests include machine learning and its application to computer vision and multimedia analysis.

Yi Yang received the Ph.D. degree in computer science from Zhejiang University. He was a Post-Doctoral Research Fellow with the School of Computer Science, Carnegie Mellon University, from 2011 to 2013. He is currently a Senior Lecturer with the Centre for Quantum Computation and Intelligent Systems, University of Technology at Sydney, Sydney. His research interests include multimedia and computer vision.

Deyu Meng (M'13) received the B.Sc., M.Sc., and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2001, 2004, and 2008, respectively. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, from 2012 to 2014. He is currently an Associate Professor with the Institute for Information and System Sciences, School of Mathematics and Statistics, Xi'an Jiaotong University. His current research interests include principal component analysis, nonlinear dimensionality reduction, feature extraction and selection, compressed sensing, and sparse machine learning methods.

Gaowen Liu received the B.S. degree in automation from Qingdao Univerisity, China, in 2006, and the M.S. degree in system engineering from the Nanjing University of Science and Technology, China, in 2008. She is currently pursuing the Ph.D. degree with the MHUG Group, University of Trento, Italy. Her research interests include machine learning and its application to computer vision and multimedia analysis.

Wei Tong received the Ph.D. degree in computer science from Michigan State University, in 2010. He is currently a Researcher with General Motors Research and Development Center. His research interests include machine learning, computer vision, and autonomous driving.

Alexander G. Hauptmann received the B.A. and M.A. degrees in psychology from Johns Hopkins University, Baltimore, MD, the degree in computer science from the Technische Universität Berlin, Berlin, Germany, in 1984, and the Ph.D. degree in computer science from Carnegie Mellon University (CMU), Pittsburgh, PA, in 1991.

He worked on speech and machine translation from 1984 to 1994, when he joined the Informedia project for digital video analysis and retrieval, and led the development and evaluation of news-on-demand applications. He is currently a Faculty Member with the Department of Computer Science and the Language Technologies Institute, CMU. His research interests include several different areas, including man–machine communication, natural language processing, speech understanding and synthesis, video analysis, and machine learning.

Nicu Sebe (M'01–SM'11) received the Ph.D. degree from Leiden University, The Netherlands, in 2001. He is currently with the Department of Information Engineering and Computer Science, University of Trento, Italy, where he leads the research in the areas of multimedia information retrieval and human behavior understanding. He is a Senior Member of the Association for Computing Machinery and a fellow of the International Association for Pattern Recognition. He was the General Co-Chair of FG 2008 and ACM Multimedia 2013, and the Program Chair of CIVR 2007 and 2010, and ACM Multimedia 2007 and 2011. He will be the Program Chair of ECCV 2016 and ICCV 2017.