Multimed Tools Appl (2018) 77:3509–3532
DOI 10.1007/s11042-017-5166-z

An efficient multi-feature SVM solver for complex event detection

Huan Liu[1] · Qinghua Zheng[1] · Zhihui Li[2] · Tao Qin[1] · Lei Zhu[3]

Received: 29 April 2017 / Revised: 7 July 2017 / Accepted: 28 August 2017/
Published online: 13 September 2017

Abstract Multimedia event detection (MED) has become one of the most important visual content analysis tools as the rapid growth of the user generated videos on the Internet. Generally, multimedia data is represented by multiple features and it is difficult to gain better performance for complex event detection with only single feature. However, how to fuse different features effectively is the crucial problem for MED with multiple features. Meanwhile, exploiting multiple features simultaneously in the large-scale scenarios always produces a heavy computational burden. To address these two issues, we propose a self-adaptive multi-feature learning framework with efficient Support Vector Machine (SVM) solver for complex event detection in this paper. Our model is able to utilize multiple features reasonably with an adaptively weighted linear combination manner, which is simple yet effective, according to the various impact that different features on a specific event. In order to mitigate the expensive computational cost, we employ a fast primal SVM solver

✉ Huan Liu
huanliucs@gmail.com

Qinghua Zheng
qhzheng@mail.xjtu.edu.cn

Zhihui Li
zhihuilics@gmail.com

Tao Qin
qin.tao@mail.xjtu.edu.cn

Lei Zhu
l.zhu@uq.edu.au

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

[2] Beijing Etrol Technologies Co., Ltd., Beijing, China

[3] School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia

in the proposed alternating optimization algorithm to obtain the approximate solution with gradient descent method. Extensive experiment results over standard datasets of TRECVID MEDTest 2013 and 2014 demonstrate the effectiveness and superiority of the proposed framework on complex event detection.

Keywords Multimedia event detection · Multi-feature learning · SVM solver

1 Introduction

Multimedia event detection (MED) has received a lot of interest largely due to the explosive growth of the user generated videos on the Internet [5, 8, 18, 42, 44]. For example, 300 hours of videos are uploaded to YouTube every minute,[1] which is the most popular video-sharing website all around the world. This task aims to identify videos of a particular event of interest, e.g., *making a cake* or *landing a fish*, which is the higher level semantic abstraction of long video clips consisting of multiple concepts [42]. For example, an event like *landing a fish* can be described by multiple concepts, such as objects (e.g., human, fish), actions (e.g., standing, pulling) and scenes (e.g., beside a river or lake). Compared with the previous visual content analysis tasks such as action detection and object recognition, MED is more challenging and complicated due to the dynamic content variations and uncontrolled capture conditions [33]. Techniques for recognizing such complex events are fundamental to many practical applications such as web video search, consumer video management, and user recommendation.

Multimedia data is usually represented by multiple features. Generally, these features can be divided into two categories, namely high-level and low-level features. Low-level features capture the local appearance and texture statistics of objects in the video at particular interest points, while high-level features are represented by a real number estimating the probability of observing a concept in the video [19]. Different features characterize different aspects of the multimedia data. Although high-performance feature descriptors have been developed to help characterize videos, it is still difficult to obtain enough required information with a single feature to discriminate between different kinds of complex events. Therefore, by common consent, combining multiple types of features or video sources is able to achieve better performance [7, 18, 19, 26, 32–34, 44]. For example, in [7] Chang et al. proposed to investigate the varying contribution of semantic representations from different image/video sources, thus enhancing the exploitation of semantic representation in the source-level. Ma et al. leveraged attributes from multiple sources to evaluate the negativity of the negative examples, demonstrating better performance than the approach that exploited each attribute source separately [26].

For a multimedia data, its multiple features, which are diverse and complementary, might assist the detection of a specific complex event on this data in varying extent. To illustrate this point more clearly, we take several different features associated with two event-related videos for an example in Fig. 1, where the top and bottom videos are with respect to the event *landing a fish* and *birthday party*, respectively. We can see that motion features such as MoSIFT [9] could be beneficial to identify the event *landing a fish*, as the event-related quick actions like "reeling or lifting" easily appear in the top video. However, for the detection of another event *birthday party*, motion features are relatively insignificant because

[1] https://www.youtube.com/yt/press/statistics.html.

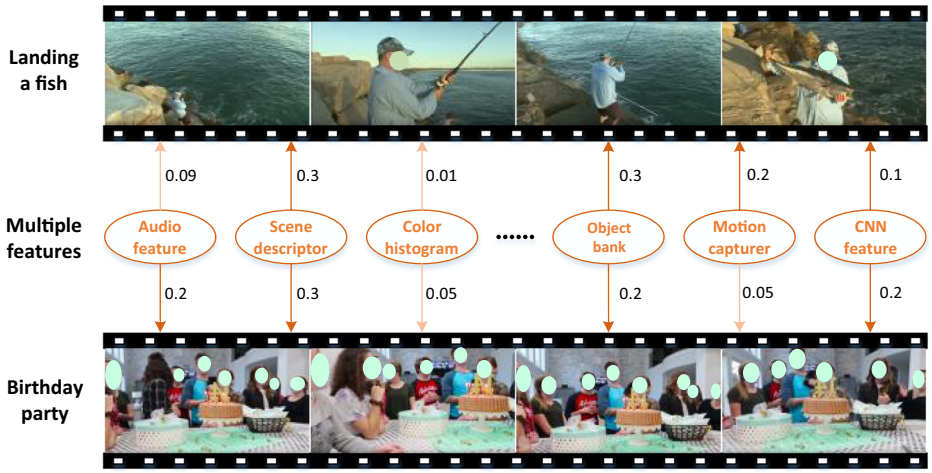


Fig. 1 An example showing the varying influence of different features with respect to the detection of a specific event. The line with dark color denotes an important impact on the event detection by the feature, while the light one means the feature is relatively insignificant to indicate the target event

these features are not the valuable indicators for this event. Different from motion features, some concepts such as scene descriptors are useful for both of these two types events. For example, "beside a river or lake" is helpful to indicate a *landing a fish* event, while "inside a room" is an important part of the event *birthday party*. On the other hand, the multi-feature representation of multimedia data such as videos usually produces high-dimension and large-volume data object. As one of the most widely used classification tools in MED community, Support Vector Machine (SVM) has been proven to be robust and effective for complex event detection task. However, previous research [2, 3, 23, 33, 41] mainly paid attention to feature selection or construction, and there are rarely studies on the problem of efficiency when applying SVM for complex event detection. As a result, multimedia event detectors have to suffer from a heavy computational burden and are time-consuming for real-world applications. In summary, we can conclude that it is challenging and complicated to involve multiple features of video data to enhance the performance of complex event detection.

In light of this, there are mainly two issues to be considered with respect to employing multiple features to enhance the performance of complex event detection. The first issue is how to leverage these diverse and complementary features reasonably when combining multiple features to detect events. The second issue is how to solve the SVM quickly to reduce the heavy computational cost, which is caused by the high-dimension and large-volume multi-feature representation. To address the both issues, we propose a self-adaptive multi-feature learning framework with the fast SVM solver for complex event detection, which is able to combine different features effectively and efficiently. In order to utilize multiple features more reasonably, we adopt an adaptively weighted linear combination for these features. This manner is simple yet effective, and is able to assign particular significance to each feature to improve the performance of MED task. Moreover, motivated by the inspiring progress in SVM-related research field [17, 25, 30], we design a fast SVM solver to alleviate the problem of expensive computational cost for complex event detection. Specifically, in the proposed alternating optimization algorithm, our solver employs an approximate

solution obtained by the gradient descent method rather than the relatively costly closed form expression. Figure 2 displays the working flow of the proposed multi-feature learning framework with efficient SVM solver for complex event detection.

We summarize our contributions as follows:

- In consideration of that different features have varying influence to indicate a specific event, we design an adaptively weighted combination manner for multiple features rather than fusing them directly to enhance the performance of complex event detection.
- In order to alleviate the heavy computational burden caused by the large-volume and high-dimension multi-feature data, the proposed alternating optimization algorithm employs an approximate solution with the gradient descent method in the large-scale scenario.
- We conduct extensive experiments on the datasets of TRECVID MEDTest 2013 and 2014 for evaluation. The promising results demonstrate the effectiveness and superiority of the proposed method.

The rest of this paper is organized as follows. In Section 2, we review related work on MED with multi-feature learning and fast primal SVM solver. Sections 3 and 4 present the details of proposed multi-feature learning framework and alternating optimization algorithm, respectively. The experimental settings and evaluation results are presented in Section 5. Section 6 concludes the paper.

2 Related work

With the rapid growth of web videos, how to exploit multiple features for complex event detection efficiently and effectively has been receiving increasing attention in recent years. We briefly review the existing related work from multiple feature learning and SVM for MED.

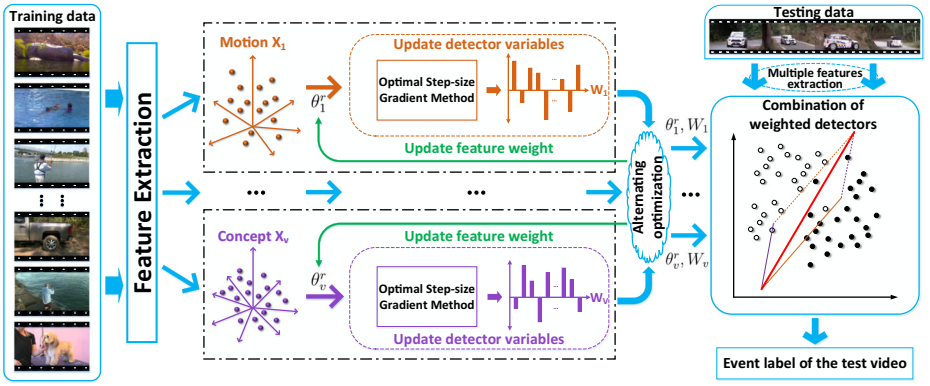


Fig. 2 The working flow of the proposed framework for complex event detection. First, our model extracts multiple features from the training data. Next, for each kind of feature, the alternating optimization is employed to update the specific weight θ_v^r and the detector W_v, which is obtained by a gradient descent method, with respect to this feature. Finally, we recognize events from testing videos by using a combination of a series of weights and detectors over multiple features

2.1 Multi-feature learning

Utilizing multiple types of features is able to achieve better performance for MED task because representing video data with a single view is rarely possible to get all required information related to the target event within such complex data. Generally, there are two major categories of multiple feature combination methods [31]. The first one is early fusion, which combines different features before the training process and then put the aggregative vector into the classifier. Spatial pyramid matching methods like [24] and [1] are the most representative research of early fusion. They have the ability to encode the spatial information of the image by fusing the features, which are extracted from different tiles generated by dividing an image. The second one is late fusion, which combines the predictive values after the training process. Some research [21, 43] have concluded that early fusion of features is less effective than late fusion in multiple content analysis when the features are independent or heterogeneous.

In light of this, plenty of research is dedicated to the study for the late fusion mechanism of multiple features. Canonical Correlation Analysis (CCA) [29], SVM-2K [13] and Multiple Kernel Learning (MKL) [16] are the most classical late fusion approaches. CCA maximizes the correlation between two features in a compact subspace. In SVM-2K, authors propose a method that combines two stage learning (kernel CCA followed by SVM) into a single optimization. MKL is widely used in computer vision but its computation is costly as the construction of multiple kernels. Recently, a number of important low-level visual features and their combination are evaluated for the complex event detection [33], which systematically analyzes these mainstream features. Yang et al. propose a semi-supervised framework [44] to improve the performance of multimedia semantic understanding by exploiting the unlabeled multiple data. In [34], Tang et al. present a method which is able to be selective of different subsets of features to combine for certain classes. Jiang et al. use a graph based approach in [19] to diffuse scores among different video data, which makes the fusion result is interpretable for human. In [6] Chang et al. present a multiple feature learning method which embedded feature interaction into a joint framework to capture the nonlinear property within the data while simultaneously combine the linear effect and the nonlinear effect. An unsupervised event saliency revealing framework which extracted features from multiple modalities is designed in [47] to represent each shot in the given video collection. Coşar et al. [11] propose a unified multimodal fusion framework that incorporates the output of object trajectory analysis with pixel-based analysis to detect abnormal behaviors related to speed and direction of object trajectories. As one of the most effectively methods for MED task, multiple feature learning also causes a heavy computational burden so that the fast optimization algorithm is desired.

2.2 SVM for MED

SVM is widely used for classification task as a result of its strong robust performance. In the field of MED, a series of SVM based algorithms have been proven to be effective for complex event detection task in both practical application [22, 23, 41, 45] and scientific research [4, 15, 36, 37]. In term of the practical application, these methods mainly contain two processes, first they construct proper features according to the characteristics of the specific data, then they directly take traditional SVM as the classifier for the final classification task. For example, Lan et al. [23] introduce double fusion scheme, which simply combines early fusion and late fusion together to incorporate their

advantages, and then employ SVM as the classifier to detect the event of interest. Xu et al. [41] propose a discriminative video representation by leveraging deep convolutional neural networks, and then apply linear SVM over the learned features to advance event detection. In a word, it is obvious that this kind of methods emphasize feature construction or combination.

In term of the scientific research, authors focus on designing SVM-based models to enhance the performance of MED. Specifically, Gkalelis et al. [15] present a two-phase approach which combines a novel nonlinear generalized subclass discriminant analysis (GSDA) method to identify a discriminant subspace, and a linear SVM to efficiently learn the event in the derived subspace. In order to deal with the problem of limited number of positive and related event videos, Tzelepis et al. [37] extend the linear SVM with Gaussian sample uncertainty (LSVM-GSU) by assuming isotropic uncertainty into a new kernel-based algorithm (KSVM-iGSU). Furthermore, they also extend KSVM-iGSU based on the relevance degree kernel SVM (RD-KSVM) proposed in [36]. As a result, related samples can be effectively exploited as positive or negative examples with automatic weighting. Recently, Chang et al. [4] present a semantic saliency and nearly-isotonic SVM framework to detect event in long videos that may last for hours. First each shot of the event is assessed and prioritized according to their saliency scores. Next, they propose a new isotonic regularizer that is able to exploit the semantic ordering information and the resulting nearly-isotonic SVM classifier exhibits higher discriminative power. However, the research on SVM solver level in MED filed to meet the growing volume of data in the large-scale scenarios is still in its infancy.

3 The proposed methodology

In this section, we explain how to construct a self-adaptive multi-view learning framework along with a generalized SVM classifier for the MED task. Suppose we have n training data represented by V different features and denote them as $X_v = [\mathbf{x}_1^v, \mathbf{x}_2^v, \cdots, \mathbf{x}_n^v] \in \mathbb{R}^{d_v \times n}$ $(v = 1, 2, \cdots, V)$, where d_v is the feature dimension of the v-th view. Let $\mathbf{y} = [y_1, y_2, \cdots, y_n]^T \in \{-1, 1\}^{n \times 1}$ are the labels of the training data, then $y_i = 1$ if the i-th video is positive example whereas $y_i = 0$ otherwise.

Generally, the low-level features are associated with the high-level events by a prediction function f. For the v-th view of the i-th video from n samples $\mathbf{x}_i^v$, we have:

$$f_i(\mathbf{x}_i^v) = W_v^\top \mathbf{x}_i^v + b_v, \tag{1}$$

where $W_v \in \mathbb{R}^{d_v \times 1}$ is the event detector with respect to the v-th view which correlates $\mathbf{x}_i^v$ with its label y_i. b_v is the bias term that also in relation to the v-th view. In order to decide f_i, we minimize the following objective:

$$\min_{f_i} loss\left(f(\mathbf{x}_i^v), y_i\right) + \Omega(f_i), \tag{2}$$

where $loss(\cdot)$ is a loss function and $\Omega(f_i)$ is the regularization function on f_i.

Considering the fact that it has been widely used and has shown its robustness, SVM is employed in this paper for MED task. Specifically, we use the hinge loss, i.e.2, $loss(f_i, y_i) = \max(1 - f_i y_i, 0)$, as the loss function and the ℓ_2-norm as the regularization

term. For better extensibility, we generalized popular hinge loss and squared hinge loss into a p-order form. Thus, for the v-th view of all n videos, we can get:

$$\min_{W_v, b_v} \sum_{i=1}^{n} \left(1 - \left(W_v^\top \mathbf{x}_i^v + b_v\right) y_i\right)_+^p + \frac{C}{2} \|W_v\|_2^2, \tag{3}$$

where the operator $(a)_+ \overset{def}{=} \max(a, 0)$ returns the scalar a if a is non-negative, and zero otherwise. Here C is the parameter to balance the relative importance of the loss term and the regularization term. p is a constant and typically $1 \leq p \leq 2$ for being meaningful.

For an event, different views usually have different contributions to the detection task as the complementary property of multiple features to each other to some extent. Therefore, it is reasonable to assign different weights $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_V]$ to different views. A larger coefficient α_v indicates that the corresponding feature plays a more important role in generating the final detector. For ease of notation, we use $L_v = \sum_{i=1}^{n} \left(1 - (W_v^\top \mathbf{x}_i^v + b_v) y_i\right)_+^p$ to denote the loss of all the videos with respect to the v-th view. Thus, the multi-view optimization problem with the p-order loss based primal SVM for the MED is:

$$\begin{aligned} \min_{\{W_v, b_v, \theta_v\}_{v=1}^V} \quad & \sum_{v=1}^{V} \left(\theta_v^r L_v + \tfrac{C}{2} \|W_v\|_2^2\right) \\ s.t. \quad & \sum_{v=1}^{V} \theta_v = 1, \quad \theta_v \geq 0 \end{aligned} \tag{4}$$

where $r > 1$. Note that if we directly use the $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_V]$ as the weight of all features, the solution to $\boldsymbol{\theta}$ is $\theta_v = 1$ for $v = \arg\min_v \{L_v\}$ and $\theta_v = 0$ otherwise. In other words, only the best view for this event is kept. Therefore, following the strategy in [38, 40], we adopt θ_v^r instead of θ_v in the objective function (4) to weight the v-th view. With this trick, our model is able to avoid degenerating into a single-view method, which considers the best view but ignores the complementary property of multiple features for an event.

Given a testing video with V features $\mathbf{x}_t^v|_{v=1}^V$, we can compute the predicted score by summing the obtained detector W_v as well as its corresponding weight θ_v^r over each view as follows:

$$y_t = \sum_{v=1}^{V} \theta_v^r W_v^\top \mathbf{x}_t^v, \tag{5}$$

where y_t is the predicted score of the testing video. To be specific, we assign positive label to this video when $y_t > 0$, otherwise negative label when $y_t \leq 0$.

In summary, by designing a multi-view learning framework for the video data, which consists of multiple features that possess complementary property to each other, our model is able to adaptively exploit different aspects of the training data for the MED task. Moreover, our model employs the p-order based hinge loss and the ℓ_2-norm based regularization term to get a more flexible SVM classifier, which has better generalization ability.

4 Optimization algorithm

In this section, we present how to obtain the event detector. Considering the non-smoothness of hinge loss used in the objective function (4), we exploit an alternating optimization algorithm to solve the proposed challenging problem effectively. We describe the alternating algorithm for optimization problem (4) in Algorithm 1.

Algorithm 1 Alternating algorithm for problem (4)

Input: Training data $X_v \in \mathbb{R}^{d_v \times n}$, $v = 1, 2, \cdots, V$, labels $\mathbf{y} \in \{-1, 1\}^{n \times 1}$, parameters $r > 1$.

Set $t = 0$, initialize $W_v \in \mathbb{R}^{d_v \times 1}$ randomly and $b_v = 0$, $\theta_v = 1/V$ for $v = 1, 2, \cdots, V$.

1: **repeat**
2: Update the feature weight coefficient θ_v^{t+1} ($v = 1, 2, \cdots, V$) with (8);
3: **for** $v = 1, \cdots, V$ **do**
4: Update event detector W_v^{t+1} for each view with Algorithm 2;
5: **end for**
6: $t = t + 1$
7: **until** Converge

Output: W_v and θ_v, $v = 1, 2, \cdots, V$.

Update feature weight First we fix $\{W_v, b_v\}|_{v=1}^V$ to update $\boldsymbol{\theta}$. In order to turn the objective function (4) into an unconstrained optimization problem, we introduce a Lagrange multiplier λ so that we have the Lagrange function of (4) as follow:

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) = \sum_{v=1}^V \left(\theta_v^r L_v + \frac{C}{2}\|W_v\|_2^2\right) - \lambda\left(\sum_{v=1}^V \theta_v - 1\right). \tag{6}$$

By setting the derivative of $\mathcal{L}(\boldsymbol{\theta}, \lambda)$ with respect to θ_v and λ to zero, we have:

$$\begin{cases} \dfrac{\partial \mathcal{L}(\boldsymbol{\theta}, \lambda)}{\partial \theta_v} = r\theta_v^{r-1} L_v - \lambda = 0, \ v = 1, 2, \cdots, V \\ \dfrac{\partial \mathcal{L}(\boldsymbol{\theta}, \lambda)}{\partial \lambda} = \displaystyle\sum_{v=1}^V \theta_v - 1 = 0. \end{cases} \tag{7}$$

Thus, θ_v can be obtained

$$\theta_v = \frac{\left(\frac{1}{L_v}\right)^{\frac{1}{r-1}}}{\displaystyle\sum_{v=1}^V \left(\frac{1}{L_v}\right)^{\frac{1}{r-1}}}. \tag{8}$$

As mentioned in Section 3, we set $r > 1$ to make the weight θ_v negatively correlate with the loss L_v. As a result, the larger the loss L_v is, the smaller the weight θ_v is. if $r \to \infty$, different weight θ_v with respect to different view will close to each other, which means all features play the same role in detecting the event. If $r \to 1$, the optimal solution to θ is $\theta_v = 1$ for $v = \arg\min_v\{L_v\}$ and $\theta_v = 0$ otherwise, which indicates only the best feature for the event is kept. Generally, the value of r should be determined according to the complementary property of all views. The view which possesses rich complementary prefers large r; otherwise, small r is suitable.

Update primal SVM variables for each feature With the feature weight $\boldsymbol{\theta}$ fixed, we update $\{W_v, b_v\}|_{v=1}^V$ with the objective function below:

$$\min_{\{W_v, b_v\}_{v=1}^V} \ \sum_{v=1}^V \left(\theta_v^r L_v + \frac{C}{2}\|W_v\|_2^2\right) \tag{9}$$

which can be decomposed into V independent problems with respect to W_v, b_v:

$$\begin{aligned} &\min_{W_v, b_v} \quad \theta_v^r L_v + \frac{C}{2}\|W_v\|_2^2 \\ \Leftrightarrow &\min_{W_v, b_v} \quad \theta_v^r \sum_{i=1}^n \left(1 - \left(W_v^\top \mathbf{x}_i^v + b_v\right) y_i\right)_+^p + \frac{C}{2}\|W_v\|_2^2 \end{aligned} \tag{10}$$

Note that $y_i \in \{-1, +1\}$ in this paper, then it is true that $1 - (W_v^\top \mathbf{x}_v^i + b_v)y_i = y_i y_i - (W_v^\top \mathbf{x}_v^i + b_v)y_i = y_i(y_i - (W_v^\top \mathbf{x}_v^i + b_v))$. Inspired by [27], we introduce the auxiliary variables $\alpha_i^v = y_i - (W_v^\top \mathbf{x}_i^v + b_v)$, $1 \le i \le n$, and the objective function with respect to W_v, b_v is equivalent to:

$$\begin{array}{ll} \min\limits_{W_v, b_v, \boldsymbol{\alpha}^v} & \theta_v^r \sum_{i=1}^n \left(y_i \alpha_i^v\right)_+^p + \frac{C}{2}\|W_v\|_2^2 \\ s.t. & \alpha_i^v = y_i - \left(W_v^\top \mathbf{x}_i^v + b_v\right),\ i = 1, 2, \cdots, n \end{array} \tag{11}$$

In order to solve the objective function (11), we adopt Augmented Lagrangian Method (ALM) [14] to update W_v, b_v and $\boldsymbol{\alpha}^v$. To begin with, by introducing a set of Lagrangian multiplier $\boldsymbol{\beta}^v$ with respect to the v-th view to take the n constraints into consideration, we have the Lagrangian function of (11) as:

$$\begin{array}{l} \mathcal{L}(W_v, b_v, \boldsymbol{\alpha}^v, \boldsymbol{\beta}^v) = \theta_v^r \sum_{i=1}^n \left(y_i \alpha_i^v\right)_+^p + \\ \frac{C}{2}\|W_v\|_2^2 + (\boldsymbol{\beta}^v)^\top \left(X_v^\top W_v + \mathbf{1}b_v - \mathbf{y} + \boldsymbol{\alpha}^v\right) \end{array} \tag{12}$$

where $\mathbf{1} = [1, 1, \cdots, 1]^\top \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{\alpha}^v = [\alpha_1^v, \alpha_2^v, \cdots, \alpha_n^v]^\top \in \mathbb{R}^{n\times 1}$. The last term is the pointwise multiplication of the amount of violation of the n constraints $\alpha_i^v - y_i + (W_v^\top \mathbf{x}_i^v + b_v) = 0$ with the vector $\boldsymbol{\beta}^v = [\beta_1^v, \beta_2^v, \cdots, \beta_n^v] \in \mathbb{R}^{n\times 1}$ consisting of n Lagrangian multipliers. Next, by adding a supplemental term to (12), we can get the augmented Lagrangian function of (11) as:

$$\begin{array}{l} \tilde{\mathcal{L}}(W_v, b_v, \boldsymbol{\alpha}^v, \boldsymbol{\beta}^v, \eta_v) = \theta_v^r \sum_{i=1}^n \left(y_i \alpha_i^v\right)_+^p + \frac{C}{2}\|W_v\|_2^2 + \\ (\boldsymbol{\beta}^v)^\top \left(X_v^\top W_v + \mathbf{1}b_v - \mathbf{y} + \boldsymbol{\alpha}^v\right) + \frac{\eta_v}{2}\|X_v^\top W_v + \mathbf{1}b_v - \mathbf{y} + \boldsymbol{\alpha}^v\|_2^2 \end{array} \tag{13}$$

where $\eta_v \in \mathbb{R}$ is the augmented penalty parameter with respect to the v-th view and "augments" to infinity and the last supplemental $\frac{\eta_v}{2}\|X_v^\top W_v + \mathbf{1}b_v - \mathbf{y} + \boldsymbol{\alpha}^v\|_2^2$ forces the n constraints to be satisfied. By arranging the last two terms in (13), we have the quadratic form as follow:

$$\begin{array}{l} \tilde{\mathcal{L}}(W_v, b_v, \boldsymbol{\alpha}^v, \boldsymbol{\beta}^v, \eta_v) = \theta_v^r \sum_{i=1}^n \left(y_i \alpha_i^v\right)_+^p + \\ \frac{C}{2}\|W_v\|_2^2 + \frac{\eta_v}{2}\|X_v^\top W_v + \mathbf{1}b_v - \mathbf{y} + \boldsymbol{\alpha}^v + \frac{\boldsymbol{\beta}^v}{\eta_v}\|_2^2 \end{array} \tag{14}$$

Compared with (13), (14) is added a term $\frac{\boldsymbol{\beta}^v}{\eta_v}$, which is always a constant when updating the variables W_v, b_v and $\boldsymbol{\alpha}^v$ within a single iteration. Note that $\eta_v \to \infty$, thus this term is almost the zero and can be negligible eventually.

After obtaining the augmented Lagrangian function (14), we then analyze how to update the event detector W_v, b_v, the auxiliary variables $\boldsymbol{\alpha}^v$, and the Lagrangian multipliers $\boldsymbol{\beta}^v$ as well as how to determine the the augmented penalty parameter η_v with respect to the v-th view. These variables will also be updated in a alternating fashion.

First we calculate the Lagrangian multiplier vector $\boldsymbol{\beta}^v$ at the t-th iteration with W_v, b_v and $\boldsymbol{\alpha}^v$ fixed. Following the strategy used in [39], we update $\boldsymbol{\beta}^v$ with the amount of violation of the n constraints:

$$\boldsymbol{\beta}_{(t)}^v = \boldsymbol{\beta}_{(t-1)}^v + (\eta_v)_{(t)} \left(X_v^\top W + \mathbf{1}b_v - \mathbf{y} + \boldsymbol{\alpha}^v\right) \tag{15}$$

Here the $(\eta_v)_{(t)}$ is monotonically non-decreasing according to the Lemma 3 in [27]. Because the first two terms $\theta_v^r \sum_{i=1}^n (y_i \alpha_i^v)_+^p$ and $\frac{C}{2}\|W_v\|_F^2$ in (14) are likely to be squeezed out by the extremely large term $\frac{\eta_v}{2}\|X_v^\top W + \mathbf{1}b_v - \mathbf{y} + \boldsymbol{\alpha}^v + \frac{\boldsymbol{\beta}^v}{\eta_v}\|_2^2$, η_v can not go to infinity in fact. Therefore, η_v can be generate under an upper bound, which is 10^5 in this paper.

Second, with W_v, b_v and $\boldsymbol{\beta}^v$ fixed, the optimization problem (14) can be decomposed into n independent problems with respect to α_i^v:

$$\begin{aligned} &\min_{\alpha_i^v} \quad \theta_v^r \left(y_i \alpha_i^v\right)_+^p + \tfrac{\eta_v}{2} \| W^\top \mathbf{x}_i^v + b_v - y_i + \alpha_i^v + \tfrac{\beta_i^v}{\eta_v} \|_2^2 \\ &\Leftrightarrow \quad \min_{\alpha_i^v} \quad \gamma_v \left(y_i \alpha_i^v\right)_+^p + \tfrac{1}{2} \left(\alpha_i^v - t_i^v\right)^2 \end{aligned} \tag{16}$$

where $\gamma_v = \frac{\theta_v^r}{\eta_v}$ and $t_i^v = y_i - W^\top \mathbf{x}_i^v - b_v - \frac{\beta_i^v}{\eta_v}$. It is easy to solve the objective function in (16) as α_i^v is the minimizer for the single-variable 2-piece piecewise function. The research [27] has given the result of a problem that is similar to ours. Based on this, for the 1-order hinge loss based primal SVM ($p = 1$), we can get:

$$\alpha_i^v = \begin{cases} t_i^v - y_i \gamma_v & \text{when } \ y_i t_i^v > \gamma_v \\ 0 & \text{when } \ 0 \le y_i t_i^v \le \gamma_v \\ t_i^v & \text{when } \ y_i t_i^v < 0 \end{cases} \tag{17}$$

Finally, the loss function term of (14) has no effect on the result when fixing $\boldsymbol{\beta}^v$ and $\boldsymbol{\alpha}^v$ to update W_v and b_v. Therefore, the optimization problem becomes:

$$\min_{W_v, b_v} \ \tfrac{C}{2} \| W_v \|_2^2 + \tfrac{\eta_v}{2} \| X_v^\top W + \mathbf{1} b_v - \mathbf{y} + \boldsymbol{\alpha}^v + \tfrac{\boldsymbol{\beta}^v}{\eta_v} \|_2^2 \tag{18}$$

Let $\boldsymbol{\tau} = \boldsymbol{\alpha}^v + \frac{\boldsymbol{\beta}^v}{\eta_v} - \mathbf{y}$ and it is easy to observe that $\boldsymbol{\tau}$ is a constant vector when update W_v and b_v. As a result, the problem (18) turns into an ℓ_2-norm regularized Least Square Regression (LSR) problem:

$$J(W_v, b_v) = \min_{W_v, b_v} \ \frac{C}{\eta_v} \| W_v \|_2^2 + \| X_v^\top W + \mathbf{1} b_v + \boldsymbol{\tau} \|_2^2 \tag{19}$$

We set $z_v = \begin{bmatrix} W_v \\ b_v \end{bmatrix}$, $A_v = \begin{bmatrix} X_v^\top & \mathbf{1} \\ (\frac{C}{\eta_v})^{\frac{1}{2}} I & \mathbf{0} \end{bmatrix}$ and $d_v = \begin{bmatrix} -\boldsymbol{\tau} \\ \mathbf{0} \end{bmatrix}$, the optimization problem (19) can be turned into a standard LSR problem as follow:

$$J(W_v, b_v) = J(z_v) = \min_{z_v} \ \| A_v z_v - d_v \|_2^2 \tag{20}$$

By setting the derivative of (20) with respect to z_v to zero, we have:

$$z_v = \left(A_v^\top A_v\right)^{-1} A_v^\top d_v. \tag{21}$$

Thus, we get the closed form of the solution with respect to W_v and b_v.

However, the time complexity of the (21) is as costly as computing matrix inverse. Considering it is time-consuming to handle the large-volume and high-dimension multi-view video data, the existing methods are not proper for such complex MED task because of the relatively high time complexity. In large-scale scenarios, usually an approximate solution of the optimization problem is enough to produce a good model [17]. Motivated by this, we

seek for an optimal step-size gradient descent method to update W_v and b_v more efficiently and effectively. The gradients of $J(W_v, b_v)$ with respect to W_v and b_v are as follows:

$$\begin{cases} W'_v = \dfrac{\partial J(W_v, b_v)}{\partial W_v} = \dfrac{C}{\eta_v} W_v + X_v \left(X_v^\top W_v + \mathbf{1} b_v + \boldsymbol{\tau} \right) \\ b'_v = \dfrac{\partial J(W_v, b_v)}{\partial b_v} = \mathbf{1}^\top \left(X_v^\top W_v + \boldsymbol{\tau} \right) + n b_v \end{cases} \tag{22}$$

Thus, we can get the optimal step-size l_v with respect to the v-th view by minimizing the single-variable quadratic function:

$$\min_{l_v} \frac{C}{\eta_v} \| W_v - l_v W'_v \|_2^2 + \| X_v^\top (W_v - l_v W'_v) + \mathbf{1}(b_v - l_v b'_v) + \boldsymbol{\tau} \|_2^2 \tag{23}$$

which has the explicit solution:

$$l_v = \frac{{W'_v}^\top W'_v + {b'_v}^2}{\left(X_v^\top W'_v + \mathbf{1} b'_v \right)^\top \left(X_v^\top W'_v + \mathbf{1} b'_v \right) + \frac{C}{\eta_v} {W'_v}^\top W'_v}. \tag{24}$$

The alternating algorithm for optimization problem (14) with respect to each feature is summarized in Algorithm 2. The time complexity of the proposed Algorithm 1 contains V parts, where V is the number of feature types. For each view, at each iteration, Algorithm 2 only needs three matrix-by-vector multiplications with complexity $\mathcal{O}(nd_v)$, where d_v is the feature dimension of the v-th view and n is the number of samples. The several pointwise addition and multiplication in (22) and (24) between two vectors are with complexity either $\mathcal{O}(d_v)$ or $\mathcal{O}(n)$, which can be neglected compared to $\mathcal{O}(nd_v)$. Therefore, the entire time complexity of the proposed Algorithm 1 is $\mathcal{O}(Vn\bar{d})$, where $\bar{d}$ is the average number of feature dimensions. In large-scale scenario, the high dimensional features are always reduced by some dimension reduction methods. At each view, the proposed Algorithm 2 has linear computational cost with respect to the number of exemplars n and is much lower than the LSQR that is as costly as computing matrix inverse.

Algorithm 2 Alternating algorithm for p-order primal SVM over single view

Input: Training data $X_v \in \mathbb{R}^{d_v \times n}$, labels $\mathbf{y} \in \{-1, 1\}^{n \times 1}$, feature weight θ_v and parameters $p = 1$, η_v.

Set $t = 0$, initialize $W_v \in \mathbb{R}^{d_v \times 1}$, $\alpha_v \in \mathbb{R}^{n \times 1}$ and $\beta_v \in \mathbb{R}^{n \times 1}$ randomly, $b_v = 0$.

1: **repeat**
2: Update the Lagrangian multiplier vector β_v^{t+1} with (15);
3: **for** $i = 1, \cdots, n$ **do**
4: Update the auxiliary variable $\alpha_v^{i\,t+1}$ with respect to each exemplar with (17);
5: **end for**
6: Calculate the gradients $W_v'^{t+1}$ and $b_v'^{t+1}$ with (22);
7: Calculate the optimal step-size l_v^{t+1} with (24);
8: Update the event detector W_v^{t+1} by $W_v^t - l_v^{t+1} W_v'^{t+1}$ and the bias term b_v^{t+1} by $b_v^t - l_v^{t+1} b_v'^{t+1}$;
9: $t = t + 1$
10: **until** Converge

Output: W_v and b_v.

Table 1 30 Events of TRECVID MEDTest 2013 and 2014

Event ID	Event Name	Event ID	Event Name
E006	Birthday party	E026	Renovating a home
E007	Changing a vehicle tire	E027	Rock climbing
E008	Flash mob gathering	E028	Town hall meeting
E009	Getting a vehicle unstuck	E029	Winning a race without a vehicle
E010	Grooming an animal	E030	Working on a metal crafts project
E011	Making a sandwich	E031	Beekeeping
E012	Parade	E032	Wedding shower
E013	Parkour	E033	Non-motorized vehicle repair
E014	Repairing an appliance	E034	Fixing musical instrument
E015	Working on a sewing project	E035	Horse riding competition
E021	Attempting a bike trick	E036	Felling a tree
E022	Cleaning an appliance	E037	Parking a vehicle
E023	Dog show	E038	Playing fetch
E024	Giving directions to a location	E039	Tailgating
E025	Marriage proposal	E040	Tuning musical instrument

5 Experiment

To demonstrate the effectiveness and superiority of the proposed framework, in this section, we conduct thorough experimental evaluation over some real-world datasets and compare with other state-of-the-art methods for complex event detection.

5.1 Datasets

We evaluate on two large scale real-world datasets: the TRECVID MEDTest 2013 [2] and the TRECVID MEDTest, 2014[3] which are collected by the NIST for the TRECVID competition. The datasets consist of about 30,000 videos from 30 events of interest, with 100 positive examples per event. Specifically, we use the videos of E006 to E015 in MEDTest 2013 dataset and videos of E021 to E040 in MEDTest 2014 dataset. Please refer to Table 1 for the complete list of event names. Several examples of the datasets used in this paper are illustrated in Fig. 3.

In order to evaluate the performance of MED with multi-feature combination, we adopt four types of features:

- SIN [28]: The SIN feature derives from the TRECVID Semantic Indexing (SIN) Task and contains 346 kinds of concepts. These concepts include objects, actions, scenes, attributes and non-visual concepts which are all the basic elements for an event, *e.g., Baby, Outdoor, Sitting down*.
- YFCC [35]: The YFCC feature derives from the Yahoo Flickr Creative Common (YFCC 100M) data which contains 0.8m Amateur videos on Flickr and 609 classes

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

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

(a) TRECVID MEDTest 2013 (b) TRECVID MEDTest 2014

Fig. 3 Example videos with respect to four events from the two datasets

of concepts. For example, the top 5 concepts detected in the YFCC100M dataset are *Outdoor, Indoor, People, Nature, Architecture*.

- SPORTS [20]: The SPORTS feature derives from The YouTube Sports-1M Dataset, which consists of 1 million YouTube videos belonging to a taxonomy of 487 classes of sports. These classes are arranged in a manually-curated hierarchical taxonomy. For example, it contains 6 different types of bowling and 7 different types of American football.
- DIY [46]. The DIY feature derives from Do it Yourself (DIY) data which is collected from online archives such as Creative Commons, Youku, Tudou and YouTube in an unsupervised fashion. These instructional videos are designed to facilitate learning for humans and include 1,601 concepts such as *Yoga, Juggling, Cooking*.

5.2 Comparison methods and experimental setup

We compare the proposed algorithm with the following important baselines:

- Standard Least Square Regression (S-LSR): This algorithm turns the optimization function into a standard LSR problem when using augmented Lagrangian method (ALM) to solve SVM problem. An exact solution can be obtained with this algorithm and the details are presented in the Section 4.
- Early Fusion (EF) [31]: EF is a combination scheme that runs before classification. We simply concatenate the four different features in a new high dimensional feature space. Disadvantage of the approach is the difficulty to combine features into a common representation. The proposed fast SVM solver is used for the classification.
- Late Fusion (LF) [31]: LF happens after classification and focuses on the individual strength of modalities. We train classifier over each feature and then combine their predictions by averaging. This scheme needs more computational effort and has potential to lose the correlation in mixed feature space. The proposed fast SVM solver is used for the classification on each feature.
- Early Fusion with Principal Component Analysis (EF-PCA): EF-PCA is a modified approach of EF based on PCA. Different from EF, We employ the PCA to reduce the dimension of the combined features. Same as EF, the proposed fast SVM solver is used for the classification.
- Rule-Based Multiple Kernel Learning (RBMKL) [12]: RBMKL is able to obtain a valid kernel by taking the summation or multiplication of several valid kernels in [12]. We use RBMKL to train an SVM with the product of the combined kernels in the experiments.

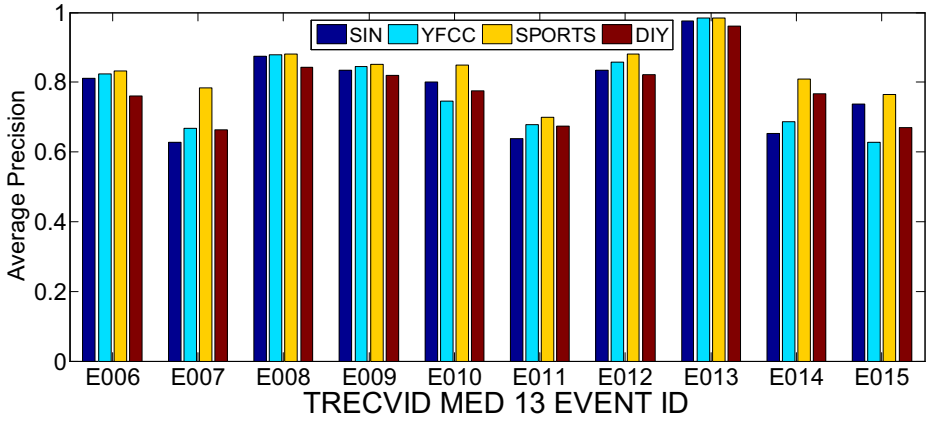


Fig. 4 Comparison of single feature with the proposed method for events E006-E015 on TRECVID MEDTest 2013

- Centered-Alignment-Based Multiple Kernel Learning (CABMKL) [10]: CABMKL is a two-stage learning algorithm. In the first step, CABMKL uses the analytical solution in [10] to determine the kernel weights. In the second step, CABMKL trains an SVM with the kernel calculated with these weights.
- The Adaptively Weighted Feature Late Fusion: The proposed algorithm which is designed for MED by using a adaptively weighted multi-feature combination manner. During each iteration, our model first updates the weights for each feature adaptively, and then applies the proposed fast SVM solver over each view for the classification.

Average precision (AP) and mean average precious (mAP) are well known and popular measures in the field of video retrieval or classification. According to the literature [42], AP is a measure combining recall and precision for ranked retrieval results. The AP is the mean of the precision scores after each relevant sample is retrieved. Generally, AP can be calculated as follows:

$$AP = \frac{1}{m} \sum_{i=1}^{n} P_i r_i \tag{25}$$

where m is the number of the relevant samples in the dataset, n is the total number of the samples, and P_i is the top-i accuracy. $r_i = 1$ when the i-th sample is relevant; otherwise $r_i = 0$. Obviously, mAP is the average performance over all events, which can be obtained with the mean all AP values. Higher value of AP (mAP) indicates better performance.

We cross-validated the regularization parameters in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}\}$, and the parameter r of the proposed method is tuned from {1.1, 5, 10, 15, 20, 25, 30}. For simplicity, we set $p = 1$ in our experiments. We report the best results for each algorithm. Particularly, we report the average of the mAP values for four features with the proposed fast SVM solver. All experiments are conducted on an 8-core Intel Xeon E5-2660 2.00 GHz Windows server with 128 G memory.

5.3 Experimental results analysis

We present the comparison of AP, mAP and training time in this section.

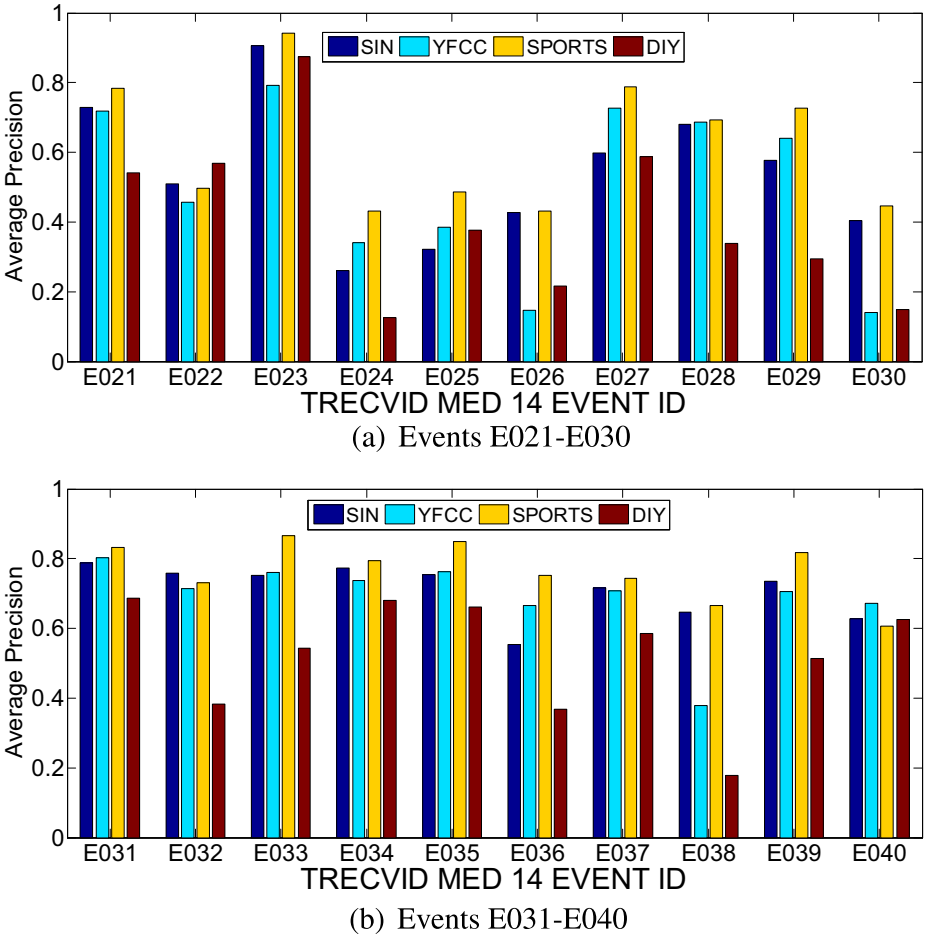


Fig. 5 Comparison of single feature with the proposed method for events E021-E030 (up) and E031-E040 (bottom) on TRECVID MEDTest 2014

AP comparison To begin with, we compare the AP performance of single feature with the proposed fast SVM solver on TRECVID MEDTest 2013 and 2014, respectively. The performance of AP with respect to each event are presented in Figs. 4 and 5. We observe from the experimental results that: 1) With the SPORTS feature, the proposed fast SVM solver achieves the best performance for 27 out of 30 events, indicating that the concepts related to some sports are useful for detection of events in TRECVID MEDTest 2013 and 2014; 2) Other features especially the SIN and YFCC have varying degree of success of getting the second place on different events, which states that giving different weights to different features is a promising fusion strategy; 3) The DIY feature is able to get better performance when detecting some instructional events such as "E007: Changing a vehicle tire", "E011: Making a sandwich", "E014: Repairing an appliance" and "E040: Tuning musical instrument."

We also present in Figs. 6 and 7 the AP values of fused features with all comparison models as well as the best single feature related to each event of TRECVID MEDTest 2013

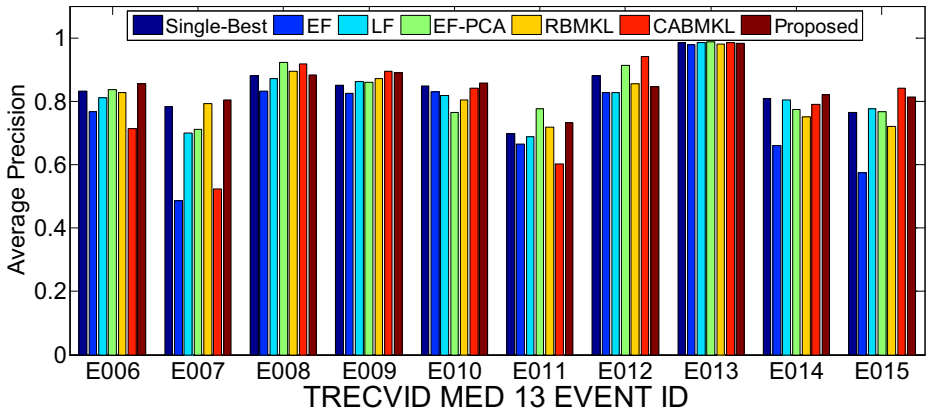


Fig. 6 Comparison of different methods of AP performance for events E006-E015 on TRECVID MEDTest 2013

and 2014, respectively. It can be concluded that: 1) The AP values of early fusion that simply concatenating different features trails behind all of other three features for 22 out of 30 events, which indicates such combination of different features is more likely to degrade the performance of classifier; 2) The kernel-based methods, i.e., RBMKL and CABMKL, are relatively effective for multiple features combination for MED except the proposed algorithm; 3) With the adaptively weighted feature fusion manner, the proposed algorithm achieves the best ot second-best performance for 25 out of 30 events. It's worth noting that our method is better than or almost equal to the best single feature. This phenomenon indicates the positive function of assigning different weights to different concepts with respect to an particular event to some extent. To be specific, for events like "E022: Cleaning an appliance" which is related to concepts contained in SIN and DIY instead of YFCC and SPORTS, our method is able to give more weights to relevant features. However, other methods treat different features equally. As a result, it is reasonable that the proposed algorithm outperforms other methods in terms of "E022: Cleaning an appliance."

mAP comparison For a fair comparison, We further compare the performance of mAP between different single feature as well as different comparison models. The values of mAP with respect to single feature over 10 events in TRECVID MEDTest 2013 and 20 events in TRECVID MEDTest 2014 with single feature are reported in Table 2 (top). The results consistently indicates that the SPORTS feature is the best among the four features on both datasets for MED task, followed by SIN and YFCC features, and DIY feature is the poorest one.

We also report in Table 2 (bottom) the values of mAP with respect to different feature fusion models over 30 events of TRECVID MEDTest 2013 and TRECVID MEDTest 2014, respectively. The experimental results indicate that: 1) The best single feature, i.e., *SPORTS*, has very good performance for MED, which shows the importance of feature construction and extraction in the field of multimedia analysis; 2) Fusing features after classification is better than before classification for the combination of multiple features for MED task. However, LF scheme still might harm the performance according to the mAP values between LF and Single(average) on TRECVID MEDTest 2014; 3) Compared with EF, EF-PCA is able to improve the mAP performance significantly because that PCA has the ability for dimension reduction and regroup of the fused features; 4) Both RBMKL

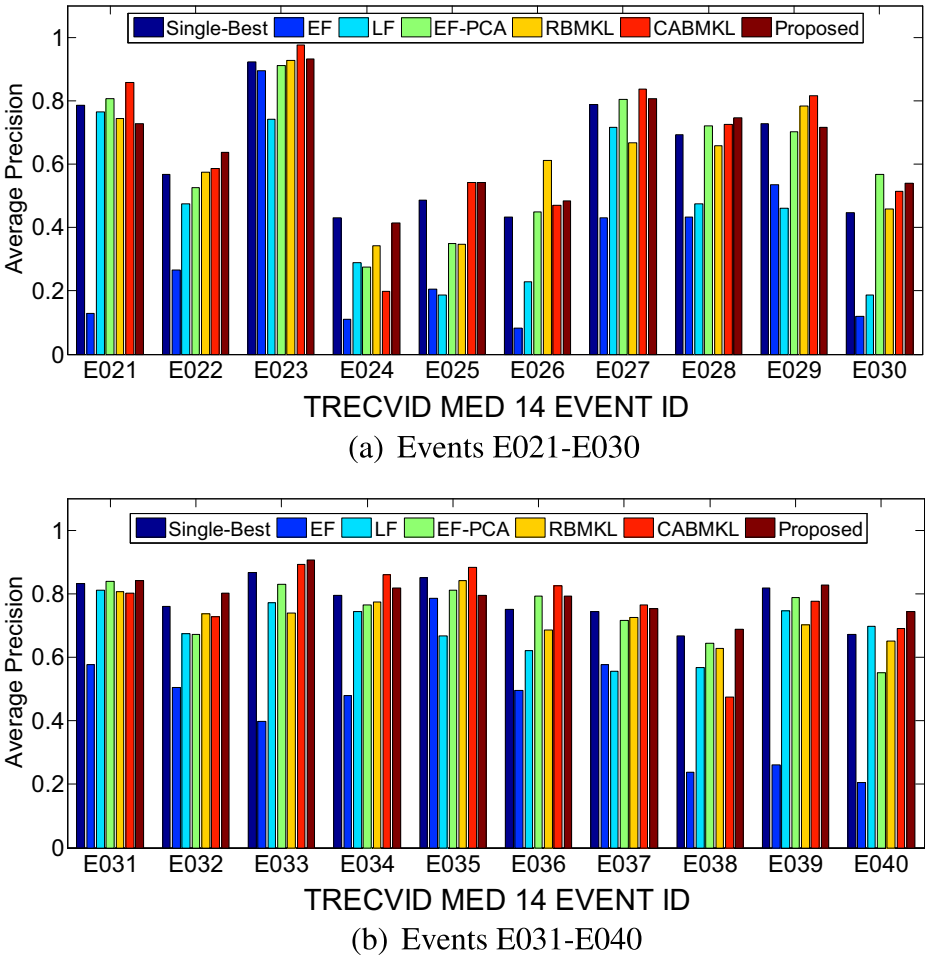


Fig. 7 Comparison of different methods of AP performance for events E021-E030 (up) and E031-E040 (bottom) on TRECVID MEDTest 2014

and CABMKL, which are based on multiple kernel learning, are competitive classifiers for MED by utilizing multiple features; 5) Our model consistently outperforms other multi-feature methods along with the best single feature with respect to both TRECVID MEDTest 2013 and 2014. Specifically, our model achieves a 16.5% on average improvement in terms of mAP comparing with the fast SVM over single feature that is widely used in MED competition. This result indicates that the proposed adaptively weighted feature fusion framework is suitable for MED with multiple features.

Training time comparison Finally, in order to evaluate the efficiency of the proposed algorithm, we compare it with the S-LSR, where the former employs the gradient descent method while the latter solves a standard LSR problem. Specifically, we calculate the training time of these two methods with the stopping condition that the variation ratio of objective is less than 10^{-4}. The results are listed in Tables 3 and 4, from which it is seen that,

Table 2 The mAP comparison of single feature with the proposed method (top) and mAP comparison of different methods (bottom) on TRECVID MEDTest 2013 and TRECVID MEDTest 2014

Datasets	SIN	YFCC	SPORTS	DIY	Average	
MEDTest13	0.779	0.779	**0.833**	0.776	0.792	
MEDTest14	0.626	0.597	**0.694**	0.465	0.596	
Datasets	EF	LF	EF-PCA	RBMKL	CABMKL	Proposed
MEDTest13	0.745	0.814	0.832	0.821	0.805	**0.849**
MEDTest14	0.386	0.569	0.676	0.694	0.711	**0.722**

The top bold emphasis indicate the SPORTS feature is the best for the complex event detection on TRECVID MEDTest2013 and 2014

The bottom bold emphasis means that the proposed method achieves the best performance on these two datasets

the proposed algorithm is with stable performance and on average faster than its competitor on both TRECVID MEDTest 2013 and 2014. In addition, the advantage of the proposed algorithm is more obvious for high-dimension datasets. For example, when the dimensionality of features increases from 346 (SIN) to 1601 (DIY) on TRECVID MEDTest 2013, the training time of the proposed algorithm increases by less than 2 times while the S-LSR's training time increases by more than 15 times. As a result, the proposed fast SVM solver is efficient for large-scale and high-dimension datasets.

Convergence analysis In order to prove the convergence of the proposed alternating optimization algorithm, we conduct some experiments on both TRECVID MEDTest 2013

Table 3 The training time (seconds) of our method and standard LSR solution over TRECVID MEDTest 2013 with different features

Event	SIN		YFCC		SPORTS		DIY	
	Proposed	S-LSR	Proposed	S-LSR	Proposed	S-LSR	Proposed	S-LSR
6	0.034	1.025	0.070	3.708	0.056	2.252	0.072	21.995
7	0.019	1.066	0.031	2.298	0.027	2.189	0.040	14.189
8	0.014	1.371	0.031	3.573	0.027	2.189	0.040	21.363
9	0.022	1.086	0.030	1.609	0.026	1.333	0.044	23.171
10	0.018	3.221	0.037	2.772	0.031	1.223	0.036	25.253
11	0.024	0.701	0.034	4.089	0.026	1.588	0.042	16.795
12	0.017	1.253	0.029	3.572	0.017	2.654	0.053	11.680
13	0.026	0.875	0.036	2.988	0.027	2.214	0.040	21.898
14	0.024	0.401	0.025	3.995	0.027	7.641	0.033	17.370
15	0.024	0.685	0.027	2.167	0.029	1.857	0.034	22.100
Average	**0.022**	1.168	**0.035**	3.077	**0.029**	2.514	**0.043**	19.582

The bold emphasis means that, in TRECVID MEDTest 2013, the proposed model is more efficient than standard LSR solution on average over different feature

Table 4 The training time (seconds) of our method and standard LSR solution over TRECVID MEDTest 2014 with different features

Event	SIN		YFCC		SPORTS		DIY	
	Proposed	S-LSR	Proposed	S-LSR	Proposed	S-LSR	Proposed	S-LSR
21	0.052	1.410	0.078	4.344	0.065	2.882	0.089	63.827
22	0.030	0.892	0.037	4.421	0.030	2.950	0.045	27.084
23	0.035	1.797	0.037	5.333	0.035	4.712	0.046	15.312
24	0.027	2.117	0.036	4.340	0.029	1.664	0.046	25.598
25	0.026	2.737	0.037	4.956	0.025	3.507	0.044	16.627
26	0.030	1.403	0.036	6.785	0.026	4.069	0.041	13.248
27	0.038	1.052	0.045	4.372	0.040	1.904	0.048	31.068
28	0.028	1.091	0.040	4.571	0.038	1.945	0.048	26.840
29	0.032	2.228	0.035	5.460	0.032	2.677	0.044	27.010
30	0.031	1.049	0.036	5.012	0.033	3.851	0.042	13.669
31	0.029	0.950	0.036	3.732	0.030	3.670	0.046	27.915
32	0.027	2.011	0.037	5.892	0.026	2.789	0.047	24.556
33	0.029	1.895	0.033	6.590	0.040	2.834	0.050	30.995
34	0.029	2.018	0.034	4.777	0.030	3.130	0.046	9.108
35	0.045	1.957	0.036	5.491	0.042	1.899	0.049	25.848
36	0.031	1.951	0.038	4.221	0.039	2.895	0.049	21.349
37	0.032	1.886	0.035	3.618	0.027	3.771	0.046	22.822
38	0.026	2.205	0.036	4.743	0.029	3.241	0.042	15.667
39	0.030	2.047	0.033	5.661	0.030	3.197	0.042	30.119
40	0.028	0.920	0.038	5.246	0.025	1.062	0.039	8.239
Average	**0.032**	1.681	**0.039**	4.978	**0.034**	2.932	**0.047**	23.845

The bold emphasis means that, in TRECVID MEDTest 2014, the proposed model is more efficient than standard LSR solution on average over different features

and 2014 to show the efficiency of our method. As the experimental results on all the 30 events are similar, we only present the convergence curves on the several events, i.e., "Birthday party", "Making a sandwich", "Attempting a bike trick" and "Beekeeping". Figure 8 shows these convergence curves. Based on our experimental results, it can be seen that the objective function value converges within 20 iterations. As a result, the convergence experiments demonstrate the efficiency of our alternating algorithm.

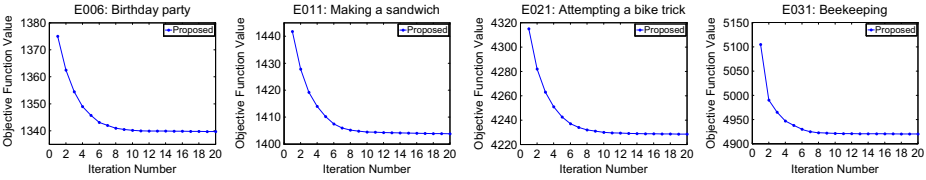


Fig. 8 Convergence curves of the objective function value in (4) using Algorithm 1 for the events "Birthday party", "Making a sandwich", "Attempting a bike trick" and "Beekeeping". The figures show that the objective function value monotonically decreases until convergence by applying the proposed algorithm

6 Conclusion

In this paper, we have explored how to fuse different features of multimedia data for MED effectively and efficiently. Considering the different impacts of different features on each event, we design an adaptively weighted combination framework for multiple features to enhance the performance of complex event detection. Moreover, in the large-scale scenario, an approximate solution with the gradient descent method is employed within the proposed alternating optimization algorithm to mitigate the heavy computational burden. We conduct extensive experiments on the datasets of MED 13 and MED 14 for evaluation. The promising results demonstrate the effectiveness and superiority of the proposed method. Finally, extensive experiment results on the datasets of TRECVID MEDTest 2013 and 2014 demonstrate the effectiveness and efficiency of the proposed method. In the future, we intend to take into account low-level features such as SIFT and MoSIFT as well as CNN features for MED except the concept-based high-level features used in this paper.

Acknowledgments This work is was supported in part by "The Fundamental Theory and Applications of Big Data with Knowledge Engineering" under the National Key Research and Development Program of China with grant Nos. 2016YFB1000903; Ministry of Education Innovation Research Team No. IRT_17R86; Project of China Knowledge Centre for Engineering Science and Technology; National Science Foundation of China under Grant Nos. 61502377.

References

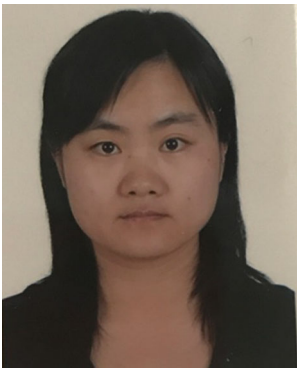
1. Bosch A, Zisserman A, Munoz X (2007) Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM international conference on image and video retrieval, pp 401–408
2. Chang X, Yang Y (2016) Semisupervised feature analysis by mining correlations among multiple tasks. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2016.2582746
3. Chang X, Nie F, Yang Y, Huang H (2014) A convex formulation for semi-supervised multi-label feature selection. In: Proceedings of the 28th AAAI conference on artificial intelligence, pp 1171–1177
4. Chang X, Yang Y, Xing EP, Yu YL (2015) Complex event detection using semantic saliency and nearly-isotonic svm. In: Proceedings of the 32nd international conference on machine learning, pp 1348–1357
5. Chang X, Yang Y, Long G, Zhang C, Hauptmann AG (2016) Dynamic concept composition for zero-example event detection. In: Proceedings of the 30th AAAI conference on artificial intelligence, pp 3464–3470
6. Chang X, Ma Z, Lin M, Yang Y, Hauptmann A (2017) Feature interaction augmented sparse learning for fast kinect motion detection. IEEE Trans Image Process 26(8):3911–3920
7. Chang X, Ma Z, Yang Y, Zeng Z, Hauptmann AG (2017) Bi-level semantic representation analysis for multimedia event detection. IEEE Trans Cybern 47(5):1180–1197
8. Chang X, Yu YL, Yang Y, Xing EP (2017) Semantic pooling for complex event analysis in untrimmed videos. IEEE Trans Pattern Anal Mach Intell 39(8):1617–1632
9. Chen MY, Hauptmann A (2009) Mosift: recognizing human actions in surveillance videos. Tech. rep. CMU-CS-09-161, Carnegie Mellon University
10. Cortes C, Mohri M, Rostamizadeh A (2010) Two-stage learning kernel algorithms. In: Proceedings of the 27th international conference on machine learning, pp 239–246
11. Coşar S, Donatiello G, Bogorny V, Garate C, Alvares LO, Brémond F (2017) Toward abnormal trajectory and event detection in video surveillance. IEEE Trans Circ Syst Vid Technol 27(3):683–695
12. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge
13. Farquhar JD, Hardoon DR, Meng H, Shawe-Taylor J, Szedmak S (2005) Two view learning: Svm-2k, theory and practice. In: Proceedings of the 19th annual conference on neural information processing systems, pp 355–362
14. Gill PE, Robinson DP (2012) A primal-dual augmented lagrangian. Comput Optim Appl 51(1):1–25

15. Gkalelis N, Mezaris V (2014) Video event detection using generalized subclass discriminant analysis and linear support vector machines. In: Proceedings of the 4th international conference on multimedia retrieval, p 25
16. Gönen M, Alpaydın E (2011) Multiple kernel learning algorithms. J Mach Learn Res 12:2211–2268
17. Hsieh CJ, Chang KW, Lin CJ, Keerthi SS, Sundararajan S (2008) A dual coordinate descent method for large-scale linear svm. In: Proceedings of the 25th international conference on machine learning, pp 408–415
18. Izadinia H, Shah M (2012) Recognizing complex events using large margin joint low-level event model. In: Proceedings of the 10th European conference on computer vision, pp 430–444
19. Jiang L, Hauptmann AG, Xiang G (2012) Leveraging high-level and low-level features for multimedia event detection. In: Proceedings of the 20th ACM international conference on multimedia, pp 449–458
20. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the 27th IEEE conference on computer vision and pattern recognition, pp 1725– 1732
21. Kludas J, Bruno E, Marchand-Maillet S (2007) Information fusion in multimedia information retrieval. In: Proceedings of the 5th international workshop on adaptive multimedia retrieval, pp 147–159
22. Lan ZZ, Jiang L, Yu SI, Rawat S, Cai Y, Gao C, Xu S, Shen H, Li X, Wang Y et al (2013) Cmu-informedia at trecvid 2013 multimedia event detection. In: Proceedings of NIST TRECVID 2013 Workshop, vol 1(2), p 5
23. Lan ZZ, Bao L, Yu SI, Liu W, Hauptmann AG (2014) Multimedia classification and event detection using double fusion. Multimed Tools Appl 71(1):333–347
24. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 19th IEEE conference on computer vision and pattern recognition, vol 2, pp 2169–2178
25. Lin CJ, Weng RC, Keerthi SS (2008) Trust region newton method for logistic regression. J Mach Learn Res 9:627–650
26. Ma Z, Chang X, Yang Y, Sebe N, Hauptmann A (2017) The many shades of negativity. IEEE Trans Multimed 19(7):1558–1568
27. Nie F, Huang Y, Wang X, Huang H (2014) New primal svm solver with linear computational cost for big data. In: Proceedings of the 31th international conference on machine learning, pp II-505
28. Over P, Fiscus J, Sanders G, Joy D, Michel M, Awad G, Smeaton A, Kraaij W, Quénot G (2014) Trecvid 2014–an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of NIST TRECVID 2014 workshop, p 52
29. Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet GR, Levy R, Vasconcelos N (2010) A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM international conference on multimedia, pp 251–260
30. Shalev-Shwartz S, Singer Y, Srebro N (2007) Pegasos: primal estimated sub-gradient solver for svm. In: Proceedings of the 24th international conference on machine learning, pp 807–814
31. Snoek CG, Worring M, Smeulders AW (2005) Early versus late fusion in semantic video analysis. In: Proceedings of the 13th annual ACM international conference on multimedia, pp 399–402
32. Song J, Yang Y, Huang Z, Shen HT, Hong R (2011) Multiple feature hashing for real-time large scale near-duplicate video retrieval. In: Proceedings of the 19th ACM international conference on multimedia, pp 423–432
33. Tamrakar A, Ali S, Yu Q, Liu J, Javed O, Divakaran A, Cheng H, Sawhney H (2012) Evaluation of low-level features and their combinations for complex event detection in open source videos. In: Proceedings of the 25th IEEE conference on computer vision and pattern recognition, pp 3681–3688
34. Tang K, Yao B, Fei-Fei L, Koller D (2013) Combining the right features for complex event recognition. In: Proceedings of the 16th IEEE international conference on computer vision, pp 2696–2703
35. Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li LJ (2015) The new data and new challenges in multimedia research. arXiv preprint arXiv:1503.01817 1(8)
36. Tzelepis C, Gkalelis N, Mezaris V, Kompatsiaris I (2013) Improving event detection using related videos and relevance degree support vector machines. In: Proceedings of the 21st ACM international conference on multimedia, pp 673–676
37. Tzelepis C, Mezaris V, Patras I (2016) Video event detection using kernel support vector machine with isotropic gaussian sample uncertainty (ksvm-igsu). In; Proceedings of the 22nd international conference on multimedia modeling, pp 3–15
38. Wang M, Hua XS, Yuan X, Song Y, Dai LR (2007) Optimizing multi-graph learning: towards a unified video annotation scheme. In: Proceedings of the 15th ACM international conference on multimedia, pp 862–871

39. Wright J, Ganesh A, Rao S, Peng Y, Ma Y (2009) Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization. In: Proceedings of the 23rd annual conference on neural information processing systems, pp 2080–2088
40. Xia T, Tao D, Mei T, Zhang Y (2010) Multiview spectral embedding. IEEE Trans Syst Man Cybern Part B (Cybern) 40(6):1438–1446
41. Xu Z, Yang Y, Hauptmann AG (2015) A discriminative cnn video representation for event detection. In: Proceedings of the 28th IEEE conference on computer vision and pattern recognition, pp 1798–1807
42. Yan Y, Yang Y, Meng D, Liu G, Tong W, Hauptmann AG, Sebe N (2015) Event oriented dictionary learning for complex event detection. IEEE Trans Image Process 24(6):1867–1878
43. Yang Y, Zhuang Y, Xu D, Pan Y, Tao D, Maybank S (2009) Retrieval based interactive cartoon synthesis via unsupervised bi-distance metric learning. In: Proceedings of the 17th ACM international conference on multimedia, pp 311–320
44. Yang Y, Song J, Huang Z, Ma Z, Sebe N, Hauptmann AG (2013) Multi-feature fusion via hierarchical regression for multimedia analysis. IEEE Trans Multimed 15(3):572–581
45. Yu SI, Xu Z, Ding D, Sze W, Vicente F, Lan Z, Cai Y, Rawat S, Schulam PF, Bahmani S et al (2012) Informedia e-lamp@ trecvid 2012: multimedia event detection and recounting (med and mer). In: Proceedings of NIST TRECVID 2012 Workshop
46. Yu SI, Jiang L, Hauptmann A (2014) Instructional videos for unsupervised harvesting and learning of action examples. In: Proceedings of the 22nd ACM international conference on multimedia, pp 825–828
47. Zhang D, Han J, Jiang L, Ye S, Chang X (2017) Revealing event saliency in unconstrained video collection. IEEE Trans Image Process 26(4):1746–1758

Huan Liu received the B.S. degree from the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2013. He is currently working toward the Ph.D. degree with the Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University. His research interests include data minning, machine learning and multimedia analysis.

Qinghua Zheng received the B.S. degree in computer software in 1990, the M.S. degree in computer organization and architecture in 1993, and the Ph.D. degree in system engineering in 1997 from Xi'an Jiaotong University, Xi'an, China. He was a Postdoctoral Researcher with Harvard University in 2002. He is currently a Professor with Xi'an Jiaotong University, and the Dean of the Department of Computer Science. His research interests include theory and technology of intelligent e-Learning environment, network public opinion and harmful information monitoring, and software reliability evaluation.



Zhihui Li received the B.S. degree from Beijing University of Posts and Telecommunications in 2008. She is currently working in Beijing Etrol Technologies Co., Ltd. Her research interests include artificial intelligence, machine learning, and computer vision.

Tao Qin received the B.S. degree in information engineering and Ph.D. degree in system engineering from the School of Electronic and Information, Xi'an Jiaotong University, Xi.an, China, in 2004 and 2010, respectively. He is currently an Associate Professor with the Systems Engineering Institute, Xi'an Jiaotong University. His research interests include Computer network measurement, network security, mobile Internet security, online social governance.

Lei Zhu received the B.S. degree (2009) at Wuhan University of Technology, the Ph.D. degree (2015) at Huazhong University of Science and Technology. He is currently a research fellow with the School of Information Technology and Electrical Engineering, University of Queensland. His research interests are in the area of large-scale image retrieval and classification.