Zhang *et al. EURASIP Journal on Image and Video Processing* (2019) 2019:44
https://doi.org/10.1186/s13640-019-0428-5

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

A feature selection framework for video semantic recognition via integrated cross-media analysis and embedded learning

Jianguang Zhang[1,2], Yahong Han[3], Jianmin Jiang[2*], Zhongrun Zhou[4], Da An[1], JieJing Liu[5] and Zhifei Song[6]

Abstract

Video data are usually represented by high dimensional features. The performance of video semantic recognition, however, may be deteriorated due to the irrelevant and redundant components included into the high dimensional representations. To improve the performance of video semantic recognition, we propose a new feature selection framework in this paper and validate it through applications of video semantic recognition. Two issues are considered in our framework. First, while those labeled videos are precious, their relevant labeled images are abundant and available in the WEB. Therefore, a supervised transfer learning is proposed to achieve the cross-media analysis, in which the discriminative features are selected by evaluating feature's correlation with the classes of videos and relevant images. Second, the labeled videos are normally rare in real-world applications. In our framework, therefore, an unsupervised subspace learning is added to retain the most valuable information and eliminate the feature redundancies by leveraging both labeled and unlabeled videos. The cross-media analysis and embedded learning are simultaneously learned in a joint framework, which enables our algorithm to utilize the common knowledge of cross-media analysis and embedded learning as supplementary information to facilitate decision making. An efficient iterative algorithm is proposed to optimize the proposed learning-based feature selection, in which convergence is guaranteed. Experiments on different databases have demonstrated the effectiveness of the proposed algorithm.

Keywords: Feature selection, Cross-media analysis, Embedded learning

1 Introduction

Video semantics recognition [1] is a fundamental research problem in computer vision [2, 3] and multimedia analysis [4, 5]. However, video data are always represented by high dimensional feature vectors [6], which often incur higher computational costs. The irrelevant and redundant features may also deteriorate the performance of video semantic recognition. In addition, feature selection [7] is able to reduce redundancy and noise information in the original feature representation, thus facilitating subsequent analysis tasks such as video semantic recognition.

Depending on whether the class label information are available, feature selection algorithms can be roughly divided into two groups [8], i.e., supervised feature selection [9] and unsupervised feature selection [10]. Supervised feature selection is able to select discriminative features by evaluating features' correlation with the classes. Thus, supervised feature selection usually yields better and more reliable performances by using the label information. However, most of the supervised feature selection methods require sufficient labeled training data in order to learn reliable model [11]. Since it is difficult to collect high-quality labeled training data in real-world applications [12], it is normally not practical to provide sufficient labeled videos for existing supervised feature selection methods to achieve satisfactory performances of feature selection. Recently, some cross-media analysis methods [13, 14] have been proposed to address the problem of insufficient number of labeled videos by transferring knowledge from other relevant types of media (e.g., images). Therefore, this type of cross-media analysis

*Correspondence: jianmin.jiang@szu.edu.cn
[2]College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China
Full list of author information is available at the end of the article

method can be considered as a kind of transfer learning. Moreover, some relevant labeled images are available and easier to collect, which can be leveraged to enhance the feature selection for video semantic recognition. To this end, we propose a supervised transfer learning in our framework, in which the knowledge from images are adapted to improve feature selection for video semantic recognition. Specifically, we use the available images with relevant semantics as our auxiliary resource and feature selection is performed on the target videos. To transfer the information from images to videos, we use the same type of still features to represent both videos and images.

Unsupervised feature selection exploits data variance and separability to evaluate feature relevance without labels. A frequently used criterion is to select the features which best preserve the data distribution or local structure derived from the whole feature set [15]. Recently, some unsupervised feature selection methods based on embedded learning have been proposed. The main advantage of utilizing embedded learning is that it can use the manifold structure of both labeled and unlabeled data to enhance the performance of feature selection. Further, most transfer learning algorithms require that the features extracted from the source domain should have the same type as that in the target domain. In practice, the videos and images in transfer learning usually need to be represented by still features such as SIFT [16]. For example, many videos are key frame-based so they cannot be represented by motion features such as STIP [17], which results in losing the underlying temporal information. To completely represent the video semantics and to effectively use the unlabeled videos, we add an unsupervised embedded learning into our proposed framework, based on augmented feature representations. To take full advantages of cross-media analysis and embedded learning, we assemble them into a joint optimization framework by introducing the joint $\ell_{2,1}$-norm regularization [18]. In this way, the information from cross-media analysis and embedded learning can be transferred from one domain to another. Moreover, the problem of over-fitting can be alleviated, and thus, the performance of feature selection can be improved. We call the proposed feature selection framework as jointing cross-media analysis and embedded learning (JCAEL). We summarize the main contributions of this paper as follows:

(1) As JCAEL can transfer the learned knowledge from relevant images to videos for improving the video feature selection, it can directly use some labeled images to address the problem of an insufficient label information. Such a merit ensures that our method is able to uncover the common discriminative features in videos and images of the same class, which provides us with better interpretability of the features.

(2) Our method contains unsupervised embedded learning, which utilizes both labeled and unlabeled videos for feature selection. This advantage guarantees that JCAEL can exploit the variance and separability of all training videos to find the common irrelevant or noisy features and thus generating optimal feature subsets. Meanwhile, videos can be represented by augmented features during the process of embedded learning, and the augmented features present more complete representation of videos, providing us the space to select the precise features of video semantics.

(3) To take the advances of cross-media analysis and embedded learning, we propose to ensemble them by adding a joint $\ell_{2,1}$-norm regularization. In this way, our algorithm is able to evaluate the informativeness of features jointly , where the correlation of features is employed. In addition, our proposed also enables cross-media analysis and embedded learning to share the common components/knowledge of features, so as to uncover common irrelevant features, which results in improving the performance of feature selection for video semantic recognition.

The rest of this paper is organized as follows. The proposed method and its corresponding optimization approach are proposed in Section 2. In Section 3, the experimental results are reported. The conclusion is shown in Section 4.

2 Proposed method

In this section, we present the framework of JCAEL. To construct this framework efficiently, we develop an iterative algorithm and prove its convergence.

2.1 Notations

To adapt knowledge from images to videos, let us denote the representations of the labeled training videos as a still feature: $X_v = \left[x_v^1, x_v^2, \ldots, x_v^{n_l}\right] \in R^{d_s \times n_l}$ where d_s is the still feature dimension and n_l is the number of the labeled training videos. Let $Y_v = \left[y_v^1, y_v^2, \ldots, y_v^{n_l}\right] \in \{0,1\}^{c_v \times n_l}$ be the labels for the labeled training videos, where c_v indicates that there are c_v different classes in videos. Similarly, we denote the representations of the images by a still feature: $X_i = \left[x_i^1, x_i^2, \ldots, x_i^{n_i}\right] \in R^{d_s \times n_i}$, where n_i is the number of the images. $Y_i = \left[y_i^1, y_i^2, \ldots, y_i^{n_i}\right] \in \{0,1\}^{c_i \times n_i}$ is the label matrix of images, where c_i indicates that there are c_i different classes in images, y_v^{kj} and y_i^{kj} denote the jth datum of y_v^k and y_i^k, $y_v^{kj} = 1$ and $y_i^{kj} = 1$ if x_v^k and x_i^k belong to the jth class; otherwise, we have $y_v^{kj} = 0$ and $y_i^{kj} = 0$. To fully utilize labeled and unlabeled videos, we use an augmented feature to denote n videos, which can be represented as $Z_v = \left[z_v^1, z_v^2, \ldots, z_v^n\right] \in R^{d_a \times n}$, where d_a is the dimension of the augmented feature. From the basic idea of feature learning, we represent the original data z_v^j

by its low dimensional embedding, i.e., $p_j \in R^{d_e}$, where d_e is the dimensionality of the embedding. As a result, the embedding of Z_v can be denoted as $P_v = \left[p_v^1, p_v^2, \ldots, p_v^n\right] \in R^{d_e \times n}$.

2.2 The proposed framework of JCAEL

We first demonstrate how to exploit the knowledge from labeled videos. To achieve this objective, learning algorithms usually use labeled training videos $\left(x_v^j, y_v^j\right)_{j=1}^{n_l}$ to learn a prediction function f that can correlate X_v with Y_v. A common approach to establish such a mechanism is to minimize the following regularized empirical error:

$$\min_f \text{loss}(f(X_v), Y_v) + \alpha\Omega(f) \tag{1}$$

where loss(.) is the loss function and $\alpha\Omega(f)$ is the regularization with α as its parameter.

It has been shown in [19] that the least square loss function gains comparable or better performance to other loss functions, such as the hinge loss, and consequently, we use the least square loss in our algorithm. The $\ell_{2,1}$-norm regularized feature selection algorithms [20, 21] utilize $\ell_{2,1}$-norm to control classifiers' capacity and also ensure there are sparse in rows, making $\ell_{2,1}$-norm particularly suitable for feature selection. Therefore, we use the $\ell_{2,1}$-norm to define the regularization, and thus, Eq. (1) can be written as

$$\min_{W_v} \left\| {W_v}^T X_v - Y_v \right\|_F^2 + \alpha \left\| W_v \right\|_{2,1} \tag{2}$$

where $W_v \in R^{d_s \times c_v}$ is the transformation matrix of the labeled videos with respect to the still feature, and $\|.\|_F$ denotes the Frobenius norm of a matrix. α is the regularization parameter. As indicated in [1, 22], the $\ell_{2,1}$-norm of W_v is defined as $\left\| W_v \right\|_{2,1} = \sum_{j=1}^{d_s} \sqrt{\sum_{k=1}^{c_v} \left(W_v^{jk}\right)^2}$, where W_v^{jk} is the jth row and the kth column element of W_v. When minimizing the $\ell_{2,1}$-norm of W_v, some rows of W_v shrink to zero, making W_v particularly suitable for feature selection.

Now, we show how to exploit the knowledge from labeled images. The fundamental step is to obtain the correlation between the images X_i and labels Y_i. Similar to Eq. (2), we achieve that by the following objective function:

$$\min_{W_i} \left\| {W_i}^T X_i - Y_i \right\|_F^2 + \alpha \left\| W_i \right\|_{2,1} \tag{3}$$

where $W_i \in R^{d_s \times c_i}$ is the transformation matrix of labeled images with respect to the still feature. When the images and videos share relevant knowledge, we can learn some shared components. Taking the semantics "playing violin" as an example, we may learn shared components about the object "violin," human action "playing," and human appearance from both videos and images. To adapt the shared information of feature selection from images to videos, we propose $\left\| W \right\|_{2,1}$ to uncover the common information shared by W_v and W_i, where $W = [W_v, W_i]$. By minimizing $\left\| W \right\|_{2,1}$, we can get sparse rows of W and uncover the common irrelevant or noisy components in both W_v and W_i. To this end, we propose the following objective function:

$$\begin{aligned} \min_{W_v, W_i} \left\| {W_v}^T X_v - Y_v \right\|_F^2 + \alpha \left\| W_v \right\|_{2,1} + \left\| {W_i}^T X_i - Y_i \right\|_F^2 \\ + \alpha \left\| W_i \right\|_{2,1} + \lambda \left\| W \right\|_{2,1} \end{aligned} \tag{4}$$

where λ is the regularization parameter.

To fully exploit both the labeled and the unlabeled videos with respect to the augmented feature representation, we show how to add the unsupervised subspace learning into Eq. (4). As it has been shown in [23] that the graph Laplacian performs well in unsupervised feature learning, we use graph Laplacian to characterize the manifold structure among the labeled and unlabeled videos. We first construct the similarity matrix S, where for the ith point z_v^i, its weight can be determined as: $S_{ij} = exp\left(-\frac{\left\| z_v^i - z_v^j \right\|_2^2}{\delta}\right)$ if and only if $z_v^j \in \mathcal{N}\left(z_v^i\right)$ or $z_v^i \in \mathcal{N}\left(z_v^j\right)$, where δ is the width parameter and $\mathcal{N}\left(z_v^t\right)$ is the k-nearest neighborhood set of z_v^t. Otherwise, $S_{ij} = 0$. As a result, the unsupervised subspace learning can be described as:

$$\begin{aligned} &\underset{P_v P_v^T = I_{d_e \times d_e}, W_z}{\arg\min} \sum_{i=1}^{n} \left\| p_v^i - \sum_{j=1}^{n} S_{ij} p_v^j \right\|_2^2 + \sum_{i=1}^{n} \left\| W_z^T z_v^i - p_v^i \right\|_2^2 \\ &= \underset{P_v P_v^T = I_{d_e \times d_e}, W_z}{\arg\min} tr\left(P_v L P_v^T\right) + \left\| W_z^T Z_v - P_v \right\|_F^2 \end{aligned} \tag{5}$$

where $p_v^i \in R^{d_e}$ is the low dimensional embedding of the original data z_v^i, d_e is the dimensionality of the embedding, $I_{d_e \times d_e}$ is the identity matrix, $W_z \in R^{d_a \times d_e}$ is the transformation matrix of videos with respect to the augmented feature, $L = (I_{n \times n} - S)^T (I_{n \times n} - S)$ is the graph Laplacian, and $P_v = \left[p_v^1, p_v^2, \ldots, p_v^n\right]$ and $tr(.)$ represent the trace operator. In Eq. (5), the most valuable information is retained and the feature redundancies are eliminated by using the low dimensional embedding p_v^i to represent the original data z_v^i. To achieve the feature selection, we use $\left\| W_z \right\|_{2,1}$ as the regularization term of Eq. (5). Therefore, the feature selection for unsupervised subspace learning can be written as:

$$\underset{P_v P_v^T = I_{d_e \times d_e}, W_z}{\arg\min} tr\left(P_v L P_v^T\right) + \left\| W_z^T Z_v - P_v \right\|_F^2 + \alpha \left\| W_z \right\|_{2,1} \tag{6}$$

As the augmented feature is the combination of still feature and motion feature, the still feature representation is a part of augmented feature representation. Since the still feature representation doesnt have motion features, we set $X_v = [X_v; 0] \in R^{d_a \times n_l}$ and $X_i = [X_i; 0] \in R^{d_a \times n_i}$, so that W_v and W_i does not affect the loss on W_z. In addition, we set $W = [W_v, W_i, W_z]$ and integrate the unsupervised subspace learning in Eq. (6) into the knowledge adaptation in Eq. (4). Finally, we arrive at the whole framework of JCAEL as follows:

$$\begin{aligned}\min_{P_v P_v^T = I_{d_e \times d_e}, W_i, W_v, W_z} & \left\| {W_v}^T X_v - Y_v \right\|_F^2 + \alpha \left\| W_v \right\|_{2,1} \\ & + \left\| {W_i}^T X_i - Y_i \right\|_F^2 + \alpha \left\| W_i \right\|_{2,1} \\ & + tr\left(P_v L P_v^T\right) + \left\| W_z^T Z_v - P_v \right\|_F^2 \\ & + \alpha \left\| W_z \right\|_{2,1} + \lambda \left\| W \right\|_{2,1}\end{aligned} \tag{7}$$

In Eq. (7), with the term $\|W\|_{2,1}$, our algorithm is able to evaluate the informativeness of the features jointly for both knowledge adaptation and low dimensional embedding. Our algorithm further enables different feature selection functions to share the common components/knowledge across knowledge adaptation and low dimensional embedding. In this way, the information from knowledge adaptation and low dimensional embedding can be transferred from one domain to the other. On the other hand, $\|W\|_{2,1}$ enables W_v, W_i, and W_z to have the same sparse patterns and share the common components, which can result in an optimal W for feature selection. Since there are four parameters (i.e., W_i, W_z, P_v, and W_v) to be estimated in Eq. (7), the objective function in Eq. (7) is not jointly convex with respect to the four parameters, but it is convex with respect to one parameter when we fix the other parameters. Thus, we propose an alternating optimization algorithm [24] to solve the optimization problem of JCAEL.

2.3 Optimization

In this section, we introduce an optimization algorithm for the objective function in Eq. (7). As there exist a number of variables to be estimated, we propose an alternating optimization algorithm to solve the optimization problem in Eq. (7). Denote $W_v = \left[w_v^1; w_v^2; \ldots w_v^{d_a}\right]$, $W_i = \left[w_i^1; w_i^2; \ldots w_i^{d_a}\right]$, $W_z = \left[w_z^1; w_z^2; \ldots w_z^{d_a}\right]$, and $W = \left[w^1; w^2; \ldots w^{d_a}\right]$, where d_a is the number of features.

(1) By fixing W_i, W_z, P_v, and optimizing W_v, the objective function in Eq. (7) can be rewritten as:

$$\min_{W_v} \left\| {W_v}^T X_v - Y_v \right\|_F^2 + \alpha \left\| W_v \right\|_{2,1} + \lambda \left\| W \right\|_{2,1} \tag{8}$$

According to [25], Eq. (8) is equivalent to

$$\min_{W_v} \left\| {W_v}^T X_v - Y_v \right\|_F^2 + \alpha tr\left(W_v^T D_v W_v\right) + \lambda tr\left(W_v^T D W_v\right) \tag{9}$$

where D_v and D are diagonal matrices with each element on the diagonal, i.e., d_v^{kk} and d^{kk} $(k = 1, 2, \ldots, d_a)$, are respectively defined as $d_v^{kk} = \frac{1}{2\left\|w_v^k\right\|_2}$ and $d^{kk} = \frac{1}{2\left\|w^k\right\|_2}$. By setting the derivative of Eq. (9) w.r.t. W_v to 0, we have

$$2X_v X_v^T W_v - 2X_v Y_v^T + 2\alpha D_v W_v + 2\lambda D W_v = 0 \tag{10}$$

Therefore, W_v can be derived by:

$$W_v = \left(X_v X_v^T + \alpha D_v + \lambda D\right)^{-1} X_v Y_v^T \tag{11}$$

(2) Similarly, by fixing W_v, W_z, P_v, and optimizing W_i, the objective function in Eq. (7) can be rewritten as:

$$\min_{W_i} \left\| {W_i}^T X_i - Y_i \right\|_F^2 + \alpha \left\| W_i \right\|_{2,1} + \lambda \left\| W \right\|_{2,1} \tag{12}$$

Similar to Eq. (8), we first denote D_i as a diagonal matrix with each element on the diagonal, i.e., d_i^{kk} $(k = 1, 2, \ldots, d_a)$, is defined as $d_i^{kk} = \frac{1}{2\left\|w_i^k\right\|_2}$. Then, Eq. (12) can be rewritten as

$$\min_{W_i} \left\| {W_i}^T X_i - Y_i \right\|_F^2 + \alpha tr\left(W_i^T D_i W_i\right) + \lambda tr\left(W_i^T D W_i\right) \tag{13}$$

By setting the derivative of Eq. (13) w.r.t. W_i to 0, we have

$$2X_i X_i^T W_i - 2X_i Y_i^T + 2\alpha D_i W_i + 2\lambda D W_i = 0 \tag{14}$$

Therefore, W_i can be optimally determined as:

$$W_i = \left(X_i X_i^T + \alpha D_i + \lambda D\right)^{-1} X_i Y_i^T \tag{15}$$

(3) By fixing W_v, W_i, P_v, and optimizing W_z, the objective function in Eq. (7) can be rewritten as:

$$\min_{W_z} \left\| W_z^T Z_v - P_v \right\|_F^2 + \alpha \left\| W_z \right\|_{2,1} + \lambda \left\| W \right\|_{2,1} \tag{16}$$

Similar to Eq. (8), we first denote D_z as a diagonal matrix with each element on the diagonal, i.e., d_z^{kk} $(k = 1, 2, \ldots, d_a)$, is defined as $d_z^{kk} = \frac{1}{2\left\|w_z^k\right\|_2}$. Then, Eq. (16) can be rewritten as

$$\min_{W_z} \left\| {W_z}^T Z_v - P_v \right\|_F^2 + \alpha tr\left(W_z^T D_z W_z\right) + \lambda tr\left(W_z^T D W_z\right) \tag{17}$$

By setting the derivative of Eq. (17) w.r.t. W_z to 0, we have

$$2Z_v Z_v^T W_z - 2Z_v P_v^T + 2\alpha D_z W_z + 2\lambda D W_z = 0 \tag{18}$$

Therefore, we have W_z to be optimally determined as:

$$W_z = \left(Z_v Z_v^T + \alpha D_z + \lambda D\right)^{-1} Z_v P_v^T \tag{19}$$

(4) By fixing W_v, W_i and substituting above W_z of Eq. (19) into Eq. (7), we will optimize P_z. Denote $A = Z_v Z_v^T + \alpha D_z + \lambda D$, the objective function in Eq. (7) can be rewritten as:

$$\min_{P_v P_v^T = I_{d_e \times d_e}} tr\left(P_v\left(L + I_{n\times n} - Z_v^T A^{-1} Z_v\right) P_v^T\right) \tag{20}$$

Considering the objective function in Eq. (20) and the constraint $P_v P_v^T = I_{d_e \times d_e}$, the optimization problem becomes

$$\begin{aligned} &\min_{P_v} tr\left(P_v\left(L + I_{n\times n} - Z_v^T A^{-1} Z_v\right) P_v^T\right) \\ &s.t.\ P_v P_v^T = I_{d_e \times d_e} \end{aligned} \tag{21}$$

If A and L are fixed, the optimization problem in Eq. (21) can be solved by Eigen-decomposition of the matrix $\left(L + I_{n\times n} - Z_v^T A^{-1} Z_v\right)$. We pick up the eigenvectors corresponding to the d_e smallest eigenvalues.

Based on the above mathematical deduction, we propose an alternating algorithm to optimize the objective function in Eq. (7), which is summarized in Algorithm 1. Once W is obtained, we sort the d_a features according to $\left\|w^k\right\|_F$ $(k = 1, 2, \ldots, d_a)$ in a descending order and select the top ranked ones.

Algorithm 1 Jointing Cross-media Analysis and Embedded Learning

Input: The labeled training videos with respect to the still feature and label information: $X_v = \left[x_v^1, x_v^2, \ldots, x_v^{n_l}\right] \in R^{d_s \times n_l}$ and $Y_v = \left[y_v^1, y_v^2, \ldots, y_v^{n_l}\right] \in \{0,1\}^{c_v \times n_l}$. The labeled training images with respect to the still feature and label information: $X_i = \left[x_i^1, x_i^2, \ldots, x_i^{n_i}\right] \in R^{d_s \times n_i}$ and $Y_i = \left[y_i^1, y_i^2, \ldots, y_i^{n_i}\right] \in \{0,1\}^{c_i \times n_i}$. The labeled and unlabeled videos with respect to the augmented feature: $Z_v = \left[z_v^1, z_v^2, \ldots, z_v^n\right] \in R^{d_a \times n}$ Parameters: α, λ, d_e.

Output: Optimized $W_v \in R^{d_a \times c_v}$, $W_i \in R^{d_a \times c_i}$ and $W_z \in R^{d_a \times d_e}$

1: Initialize W_v, W_i and W_z randomly.
2: Compute W according to $W = [W_v, W_i, W_z]$.
3: Construct the graph Laplacian matrix $L \in R^{n \times n}$.
4: **repeat**
5: Compute diagonal matrices D_v, D_i, D_z and D respectively.
6: Compute W_v according to Eq. (11).
7: Compute W_i according to Eq. (15).
8: Compute P_v according to Eq. (21).
9: Compute W_z according to Eq. (19).
10: Compute W according to $W = [W_v, W_i, W_z]$.
11: **until** Convergence
12: **return** W_v, W_i, W_z and W

2.4 Convergence and computational complexity

2.4.1 Convergence

In this section, we theoretically show that Algorithm 1 proposed in this paper converges. We begin with the following lemma [22].

Lemma 1 *For any nonzero vectors w and $\widehat{w}$, the following inequality holds:*

$$\|w\|_2 - \frac{\|w\|_2^2}{2\|\widehat{w}\|_2} \le \|\widehat{w}\|_2 - \frac{\|\widehat{w}\|_2^2}{2\|\widehat{w}\|_2} \tag{22}$$

As a result, the second lemma can be derived as described below.

Lemma 2 *By fixing W_i and W_v, we obtain the global solutions for W_z and P_v in Eq. (7). Yet, by fixing W_i, W_z, and P_v, we obtain the global solutions for W_v in Eq. (7). In the same manner, by fixing W_v, W_z, and P_v, we obtain the global solutions for W_i in Eq. (7).*

Proof When W_i and W_v are fixed, the optimization problem in Eq. (7) is equivalent to the problem described in Eq. (17) and Eq. (21). We can solve the convex optimization problem with respect to W_z by setting the derivative of (17) to zero. Further, we can derive the global solution for P_v by solving the Eigen-decomposition problem with respect to P_v. When W_z, P_v, and W_i are fixed, the optimization problem in Eq. (7) is equivalent to the problem described in Eq. (9). We can solve the convex optimization problem with respect to W_v by setting the derivative of Eq. (9) to zero. Thus, we derive the global solution for W_v in Eq. (7), provided that W_z, P_v, and W_i are fixed. Similarly, we can also derive the same conclusion when W_i is fixed. □

Theorem 1 *The proposed algorithm monotonically decreases the objective function value of Eq. (7) in each iteration. Next, we prove Theorem 1 as follows.*

Proof Let $\widehat{W}_v$, $\widehat{W}_i$, $\widehat{P}_v$, and $\widehat{W}_z$ denote the updated W_v, W_i, P_v, and W_z, respectively. The loop to update W_v, W_i, P_v, and W_z in the proposed algorithm corresponds to the optimal W_v, W_i, P_v, and W_z of the following problem:

$$\begin{aligned} \min_{P_v P_v^T = I_{d_e \times d_e}, W_i, W_v, W_z} & \left\| {W_v}^T X_v - Y_v \right\|_F^2 + \alpha \left\| W_v \right\|_{2,1} \\ & + \left\| {W_i}^T X_i - Y_i \right\|_F^2 + \alpha \left\| W_i \right\|_{2,1} \\ & + tr\left(P_v L P_v^T\right) + \left\| W_z^T Z_v - P_v \right\|_F^2 \\ & + \alpha \| W_z \|_{2,1} + \lambda \left\| W \right\|_{2,1} \end{aligned} \tag{23}$$

Since $\|W\|_{2,1} = \sum_{k=1}^{d_a} \left\|w^k\right\|_2$ [26], according to Lemma 2, we can obtain:

$$\begin{aligned}
&\left\|\widehat{W}_v^T X_v - Y_v\right\|_F^2 + \alpha\left\|\widehat{W}_v\right\|_{2,1} + \left\|\widehat{W}_i^T X_i - Y_i\right\|_F^2 + \alpha\left\|\widehat{W}_i\right\|_{2,1} + tr\left(\widehat{P}_v L \widehat{P}_v^T\right) \\
&\quad + \left\|\widehat{W}_z^T Z_v - \widehat{P}_v\right\|_F^2 + \alpha\left\|\widehat{W}_z\right\|_{2,1} + \lambda\sum_k \frac{\left\|\widehat{w}^k\right\|_2^2}{2\left\|w^k\right\|_2} \\
&\quad \le \left\|W_v{}^T X_v - Y_v\right\|_F^2 + \alpha\|W_v\|_{2,1} + \left\|W_i{}^T X_i - Y_i\right\|_F^2 + \alpha\|W_i\|_{2,1} \\
&\quad + tr\left(P_v L P_v^T\right) + \left\|W_z^T Z_v - P_v\right\|_F^2 + \alpha\|W_z\|_{2,1} + \lambda\sum_k \frac{\left\|w^k\right\|_2^2}{2\left\|w^k\right\|_2}
\end{aligned} \tag{24}$$

Then, we have the following inequality:

$$\begin{aligned}
&\left\|\widehat{W}_v^T X_v - Y_v\right\|_F^2 + \alpha\left\|\widehat{W}_v\right\|_{2,1} + \left\|\widehat{W}_i^T X_i - Y_i\right\|_F^2 + \alpha\left\|\widehat{W}_i\right\|_{2,1} \\
&\quad + tr\left(\widehat{P}_v L \widehat{P}_v^T\right) + \left\|\widehat{W}_z^T Z_v - \widehat{P}_v\right\|_F^2 + \alpha\left\|\widehat{W}_z\right\|_{2,1} \\
&\quad + \lambda\sum_k \left\|\widehat{w}^k\right\|_2 - \lambda\left(\sum_k \left\|\widehat{w}^k\right\|_2 - \sum_k \frac{\left\|\widehat{w}^k\right\|_2^2}{2\left\|w^k\right\|_2}\right) \\
&\quad \le \left\|W_v{}^T X_v - Y_v\right\|_F^2 + \alpha\|W_v\|_{2,1} + \left\|W_i{}^T X_i - Y_i\right\|_F^2 + \alpha\|W_i\|_{2,1} \\
&\quad + tr\left(P_v L P_v^T\right) + \left\|W_z^T Z_v - P_v\right\|_F^2 + \alpha\|W_z\|_{2,1} \\
&\quad + \lambda\sum_k \left\|w^k\right\|_2 - \lambda\left(\sum_k \left\|w^k\right\|_2 - \sum_k \frac{\left\|w^k\right\|_2^2}{2\left\|w^k\right\|_2}\right)
\end{aligned} \tag{25}$$

According to Lemma 1, another inequality can be established as follows:

$$\begin{aligned}
&\left\|\widehat{W}_v^T X_v - Y_v\right\|_F^2 + \alpha\left\|\widehat{W}_v\right\|_{2,1} + \left\|\widehat{W}_i^T X_i - Y_i\right\|_F^2 + \alpha\left\|\widehat{W}_i\right\|_{2,1} + tr\left(\widehat{P}_v L \widehat{P}_v^T\right) \\
&\quad + \left\|\widehat{W}_z^T Z_v - \widehat{P}_v\right\|_F^2 + \alpha\left\|\widehat{W}_z\right\|_{2,1} + \lambda\sum_k \left\|\widehat{w}^k\right\|_2 \\
&\quad \le \left\|W_v{}^T X_v - Y_v\right\|_F^2 + \alpha\|W_v\|_{2,1} + \left\|W_i{}^T X_i - Y_i\right\|_F^2 + \alpha\|W_i\|_{2,1} \\
&\quad + tr\left(P_v L P_v^T\right) + \left\|W_z^T Z_v - P_v\right\|_F^2 + \alpha\|W_z\|_{2,1} + \lambda\sum_k \left\|w^k\right\|_2
\end{aligned} \tag{26}$$

□

This indicates that, with the updating rule in the proposed algorithm, the objective function value for Eq. (7) monotonically decreases until a convergence is reached.

2.4.2 Computational complexity

For the computational complexity of Algorithm 1, computing the graph Laplacian matrix L is $O\left(n^2\right)$. During the training, learning W_v, W_i, and W_z involves calculating the inverse of a number of matrices, among which the most complex part is $O\left(d_a^3\right)$. To optimize the P_v, the most time-consuming operation is to perform eigen-decomposition of the matrix $ED = \left(L + I_{n\times n} - Z_v^T A^{-1} Z_v\right)$. Note that $ED \in R^{n\times n}$. The time complexity of this operation is $O\left(n^3\right)$ approximately. Thus, the computational complexity of JCAEL can be worked out as $max\left\{O\left(t\times n^3\right), O\left(t\times d_a^3\right)\right\}$, where t is the number of iterations required for convergence. From the experiments, we observe that the algorithm converges within 10 $\sim$ 15 iterations, which indicates that our proposed algorithm is efficient in feature selection for video semantics recognition.

3 Experimental results and discussion

In this section, we propose the video semantic recognition experiments which evaluate the performance of our jointing cross-media analysis and embedded learning (JCAEL) for feature selection.

3.1 Experimental datasets

In order to evaluate the contribution from cross-media analysis, we construct three couples of video and image datasets, which include HMDB13 (video dataset)←"Extensive Images Databases" (EID, image dataset), UCF10 (video dataset)← Actions Images Databases (AID, image dataset), UCF (video dataset)←PPMI4 (image dataset), where "←" denotes the direction of adaptation from images to videos. The videos and images of HMDB13←EID and UCF10←AID have the same semantic classes, and UCF←PPMI4 has different semantic classes for videos and images.

3.1.1 HMDB13←EID

The HMDB51 dataset [27] is collected from a variety of sources ranging from digitized movies to YouTube videos. It contains 6766 video sequences that are categorized into 51 classes. This dataset contains simple facial actions, general body movements, and human interactions. In order to increase the number of overlapping classes, we select 13 overlapping classes between HMDB51 and another image datasets as Extensive Images Databases (EID), which includes two open benchmark datasets (i.e., Stanford40 [28] and Still DB [29]). As a result, we call the video dataset as "HMDB13." Table 1 provides the details of the overlapping classes from EAD to HMDB51.

Table 1 The classes of HMDB13 ← EID

Datasets	HMDB13 (video dataset)	EID (image dataset)
The overlapping classes	Catch	Catching (Still DB)
	Clap	Applauding (Stanford40)
	Drink	Drinking (Stanford40)
	Jump	Jumping (Stanford40)
	Pour	Pouring liquid (Stanford40)
	Pushing	Pushing a cart (Stanford40)
	Run	Running (Stanford40)
	Smoke	Smoking (Stanford40)
	Wave	Waving hands (Stanford40)
	Kick	Kicking (Still DB)
	Throw	Throwing (Still DB)
	Walk	Walk (Still DB)
	Climbing	Climbing (Stanford40)

3.1.2 UCF10←AID

The UCF101 [30] is a dataset of realistic action videos collected from YouTube, which has 101 action categories. It gives the largest diversity in terms of actions with the presence of large variation in subject appearances, including scale and pose, related objects, cluttered background, and illumination conditions. Such a challenging diversity is suitable for verifying the effect of information learned from images on video semantics recognition. To further evaluate whether images coming from various sources contribute to the feature selection or not, we select ten overlapping classes between UCF101 and the action image dataset, referred to as Actions Images Databases (AID), which includes four open benchmarking datasets (i.e., action DB [31], PPMI [32], willow-actions [33], and still DB). For the convenience of our experiments design and description, we call the video dataset as UCF10. In Table 2, we show the chosen categories of UCF10←AID, which are taken as video dataset and image dataset, respectively.

3.1.3 UCF←PPMI4

The PPMI dataset [32] consists of 7 different musical instruments: bassoon, erhu, flute, French horn, guitar, saxophone, and violin. In order to assess the performance of the proposed algorithm when the image dataset has different classes from that in the video dataset, we choose ten classes from UCF101 and then select four overlapping image categories from PPMI. To this end, we call the video dataset as UCF and image dataset as PPMI4. Table 3 summarizes the selected classes of UCF←PPMI4.

3.2 Experiment setup

For all the datasets, we select 30 images from each overlapping categories for knowledge adaption as the number of images is relatively small. We sample videos for labeled training data and take the remaining videos as the testing data. To evaluate the contribution from unsupervised subspace learning, we conduct experiments to study the performance variance when only a few labeled training samples are provided, and the ratios of labeled video data are set to 5%. For each dataset, we repeat the sampling for 10 times and report the average results. We extract SIFT features [16, 34] from the key frames of videos and images. The STIP features [17] are extracted from videos. We use the standard Bag-of-Words (BoW) method [35, 36] to generate the BoW representation of SIFT and STIP features, where the number of visual words of Bag-of-Words is set to 600. For videos, we obtain a still feature with 600 dimensions and an augmented feature with 1200 dimensions, and for images, we obtain a still feature with 600 dimensions.

Table 2 The classes of UCF10←AID

Datasets	UCF10 (video dataset)	AID (image dataset)
The overlapping classes	Biking	Riding bike (willow-actions)
	Cricket bowling	Cricket bowling (action DB)
	Cricket shot	Cricket batting (action DB)
	Horse riding	Riding horse (willow-actions)
	Playing cello	Playing cello (PPMI)
	Playing flute	Playing flute (PPMI)
	Playing violin	Playing violin (PPMI)
	Tennis swing	Tennis forehand (action DB)
	Volleyball spiking	Volleyball smash (action DB)
	Base ball pitch	Throwing (still DB)

Table 3 The classes of UCF←PPMI4

Datasets	UCF (video dataset)	PPMI4 (image dataset)
The overlapping classes	Playing cello	Playing cello (PPMI)
	Playing flute	Playing flute (PPMI)
	Playing guitar	Playing guitar (PPMI)
	Playing violin	Playing violin (PPMI)
	Rock climbing in door	NULL
	Rowing	NULL
	Tennis swing	NULL
	Volleyball spiking	NULL
	Walking with dog	NULL
	Writing on board	NULL

3.3 Comparison algorithms

To benchmark our proposed jointing cross-media analysis and embedded learning (JCAEL), we select a number of representative existing state of the arts for performance comparisons, details of which are highlighted below:

- Full features (FF) which adopts all the features for classification. It is used as baseline method in this paper.
- Fisher score feature selection (FSFS) [37]: a supervised feature selection method built by depending on fully labeled training data to select features with the best discriminating ability.
- Feature selection via joint $\ell_{2,1}$-norms minimization (FSNM) [22]: a supervised feature selection method built by employing joint $\ell_{2,1}$-norms minimization on both loss function and regularization to realize feature selection across all data points.
- $\ell_{2,1}$-norm least square regression (LSR_{21}) [22]: a supervised feature selection method built upon least square regression by using the $\ell_{2,1}$-norm as the regularization term.
- Multi-class $\ell_{2,1}$-norm support vector machine (SVM_{21}) [20]: a supervised feature selection method

built upon SVM by using the $\ell_{2,1}$-norm as the regularization term.
- Ensemble feature selection (EnFS) [25]: a supervised feature selection method based on transfer learning, which transfer the shared information between different classifiers by adding a joint $\ell_{2,1}$-norm on multiple feature selection matrices.
- Joint embedding learning and sparse regression (JELSR) [26]: a unsupervised feature selection method built by using the local linear approximation weights and $\ell_{2,1}$-norm regularization.
- Jointing cross-media analysis and embedded learning (JCAEL): our proposed method which is designed for feature selection by adapting knowledge from images based on still feature and utilizing both labeled and unlabeled videos based on augmented feature.

During the process of training and predicting, we use the augmented feature to represent the videos for the baseline methods, including FSFS, FSNM, LSR_{21}, SVM_{21}, and JELSR as these methods cannot use the information adapted from images. For EnFS and JCAEL, we use the still features to represent the image data and use the augmented feature to represent the videos. To fairly compare different feature selection algorithms, we use a "grid-research" strategy from $\{10^{-6}, 10^{-5}, \ldots, 10^{5}, 10^{6}\}$ to tune the parameters for all the compared algorithms. By setting the number of selected features as $\{120, 240, \ldots, 1200\}$, we report the best results obtained from different parameters. For the K-nearest neighbors of Laplacian matrix L, the parameter is set to $k = 10$. In our experiment, each feature selection algorithm is first performed to select features. Then, three classifiers, i.e., linear multi-class SVM (LMCSVM), least square regression (LSR), and multi-class kNN(MCkNN), are performed based on the selected features respectively to assess the performance of feature selection. For the classifier of least square regression, we learn a threshold from the labeled training data to quantize the continuous label prediction scores to binary. To measure the feature selection performances, we use the average accuracy (AA) over all semantic classes as the evaluation metric, which is defined as:

$$\mathrm{AA} = \frac{\sum_{k=1}^{c_v} \mathrm{acc}_k}{c_v} \tag{27}$$

where c_v is the number of action classes. acc_k is the accuracy for the kth class.

3.4 Experimental results

In order to evaluate the effectiveness of JCAEL, we compare JCAEL with FF, FSFS, FSNM, LSR_{21}, SVM_{21}, EnFS, and JELSR on both HMDB13←EID and UCF10←AID dataset. The comparison results are summarized in Tables 4 and 5, where the best and the second best results are highlighted in bold and italic, correspondingly. We also conduct a number of experiments to study the performance variance when the ratios of labeled video data are set to 5%, 10%, 20%, 30%, and 40%, and the results are displayed in Figs. 1 and 2.

From the experimental results in Tables 4–5 and Figs. 1–2, we can make the following observations:

(1) The results of feature selection algorithms are generally better than that of full features (FF). As the classification could be much faster by reducing the feature number, feature selection proves to be more crucial in practical applications.
(2) As the number of labeled training videos increases, the performance of all methods is improved. This is consistent with the general principle as more information is made available for training.
(3) The classification using multi-class SVM and multi-class kNN achieve better performance than the least square regression when the ratio of labeled video data are set to 5%. The main reason is that the threshold learned from the small size of training data leads to a bias in the quantization of continuous label prediction scores.
(4) When the ratio of labeled video data are set to 5%, JELSR is generally the second most competitive algorithm. This indicates that incorporating the additional information contained in the unlabeled training data through unsupervised embedded learning is indeed useful.
(5) As shown in Figs. 1–2, supervised methods based on transfer learning (EnFS) always achieve better performances than other compared methods when the number of labeled training videos is enough (e.g., the ratio of labeled video data are set to 40%), since EnFS can uncover common irrelevant features by transferring the relative information between different classifiers.

Table 4 Comparisons of feature selection algorithms on HMDB13←EID in terms of average accuracy using three classifiers when the ratio of labeled video data are set to 5%

Classifiers	FF	FSFS	FSNM	LSR_{21}	SVM_{21}	EnFS	JELSR	JCAEL
LMCSVM	0.3032	0.3187	0.3137	0.3117	0.3137	0.3182	***0.3387***	**0.3526**
LSR	0.1763	0.2138	0.1898	0.1903	0.1873	0.2003	***0.2233***	**0.2318**
MCkNN	0.1898	0.2647	0.2517	0.2582	0.2093	0.2697	***0.2737***	**0.2877**

Table 5 Comparisons of feature selection algorithms on UCF10←AID in terms of average accuracy using three classifiers when the ratio of labeled video data are set to 5%

Classifiers	FF	FSFS	FSNM	LSR_{21}	SVM_{21}	EnFS	JELSR	JCAEL
LMCSVM	0.4340	0.4448	0.4715	0.4355	0.4340	0.4348	***0.5061***	**0.5299**
LSR	0.2906	0.3136	0.3043	0.3057	0.3028	0.3064	***0.3180***	**0.3302**
MCkNN	0.3360	0.3684	0.3670	0.3360	0.3360	0.3381	***0.3756***	**0.4001**

(6) As shown in Figs. 1–2, our proposed JCAEL remains to be the best performing algorithm among different methods and different cases. The main reason is that our method can take advantages of both transfer learning and embedded learning. We can also see from Tables 4–5 that JCAEL algorithm achieves the best results when only a small number of labeled training videos are available. This advantage is especially desirable for real-world problems since precisely annotated videos are often rare.

3.5 Experiment on convergence

In this section, we study the convergence of the proposed JCAEL as described in Algorithm 1. Due to the fact that we solve our objective function using an alternating approach, how fast our algorithm converges is crucial for the whole computational efficiency in practice. Hence, we conduct an experiment to test the convergence of the proposed JCAEL algorithm according to the objective function value in Eq. (7) on both HMDB13←EID and UCF10←AID datasets, where the ratio of labeled video

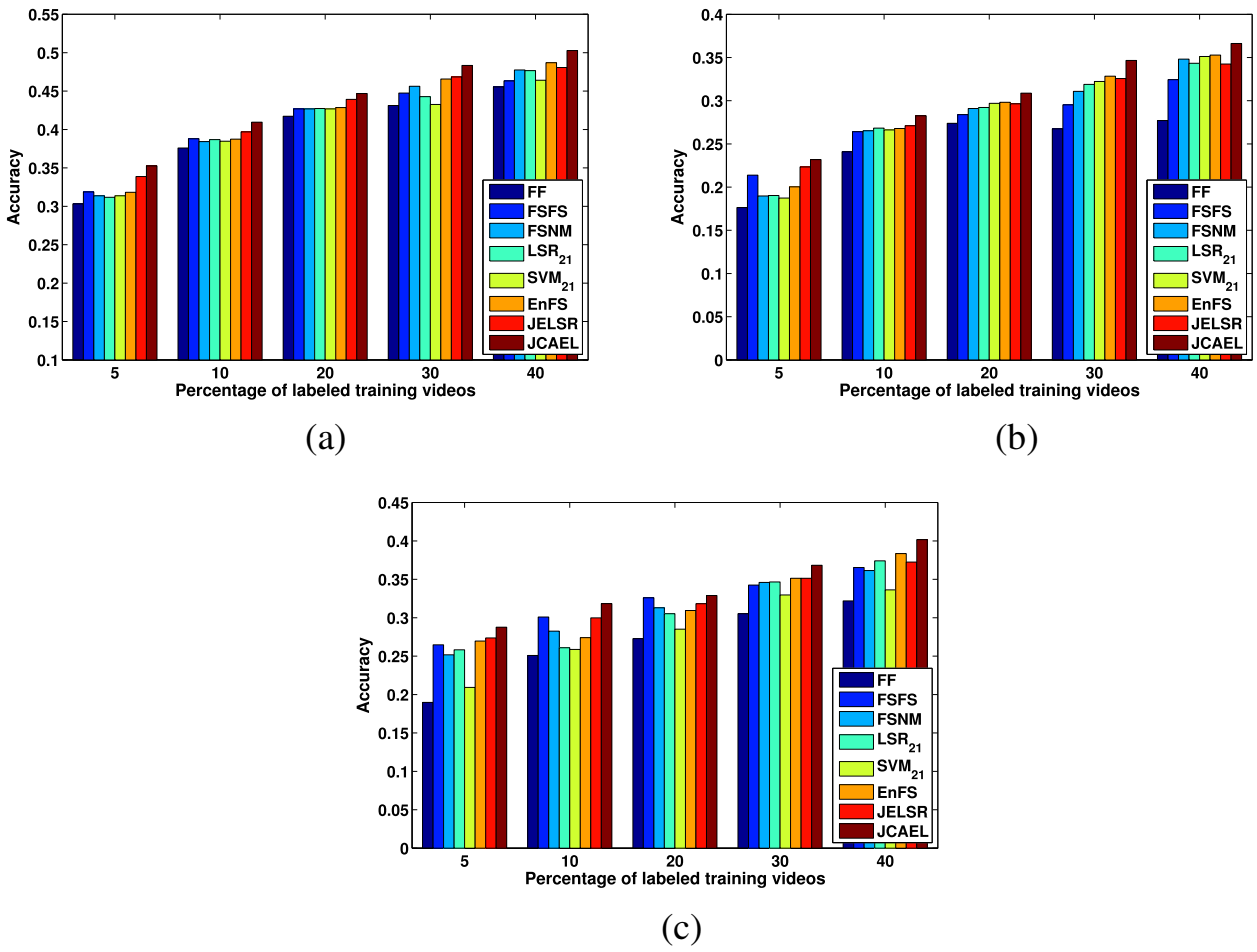

Fig. 1 Classification results of different methods on HMDB13←EID when the ratios of labeled video data are different. **a** LMCSVM as classifier. **b** LSR as classifier. **c** MCkNN as classifier

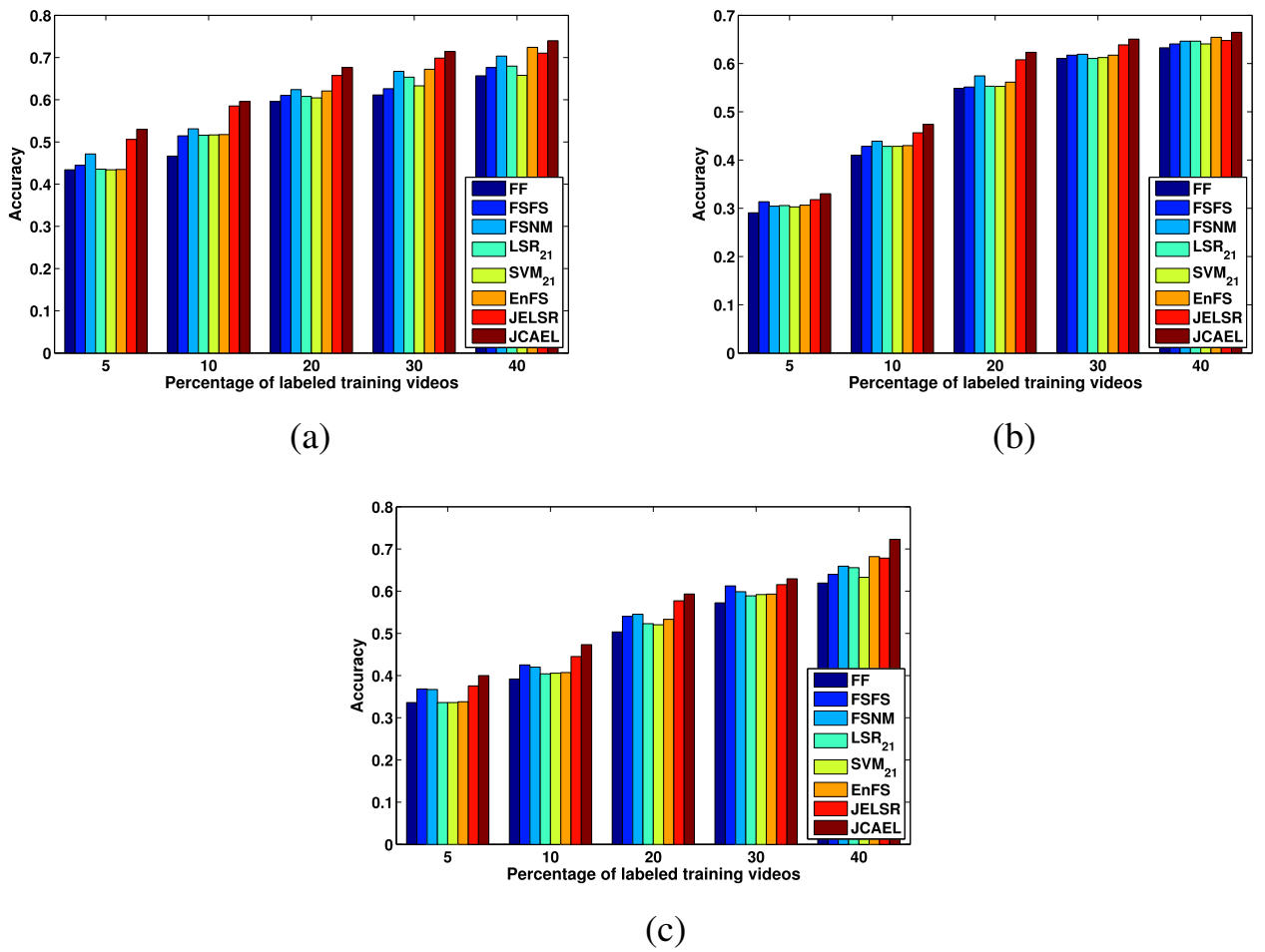

Fig. 2 Classification results of different methods on UCF10←AID when the ratios of labeled video data are different. **a** LMCSVM as classifier. **b** LSR as classifier. **c** MCkNN as classifier

data are set to 5%. All the results are illustrated as convergence curves, and when the ratio of labeled video data are set to 40%, all the results are summarized in Fig. 3, where all the parameters involved are fixed at their optimal values. From the results shown in Fig. 3, it can be seen that our algorithm converges within a few iterations. For example, it takes no more than 10 iterations for UCF10←AID and no more than 15 iterations for HMDB13←EID.

3.6 Experiment on parameter sensitivity

There are two regularization parameters α and λ in Eq. (7). To learn how they affect the performances, we conduct an experiment to test the parameter sensitivity, where LMCSVM is used to classify the videos. We show the results on both HMDB13←EID and UCF10←AID in Fig. 4, where the ratio of labeled video data are set to 5%. It can be seen that, for HMDB13←EID, the performance is sensitive to the two parameters. For UCF10←AID the performance does not change much. In general, our proposed can perform well for these datasets when α and λ are comparable. For example, good performance is obtained when $\alpha = 0.0001$ and $\lambda = 100$ for HMDB13←EID and $\alpha = 0.001$ and $\lambda = 10000$ for UCF10←AID.

3.7 Experiment on selected features

As feature selection is aimed at both accuracy and computational efficiency, we perform an experiment to study how the number of selected features can affect the performance. We construct the experiments on both HMDB13←EID and UCF10←AID when the ratio of labeled video data is set to 5%. Again, LMCSVM is used to classify the videos, and Fig. 5 shows the performance variation w.r.t the number of selected features. From the results illustrated in Fig. 5, the following observations can be made: (1) When the number of selected features is too small, the result is not competitive with using all features for video semantic recognition, which could be attributed to the fact that too much information is lost in this case. For instance, when using less than 360 features of

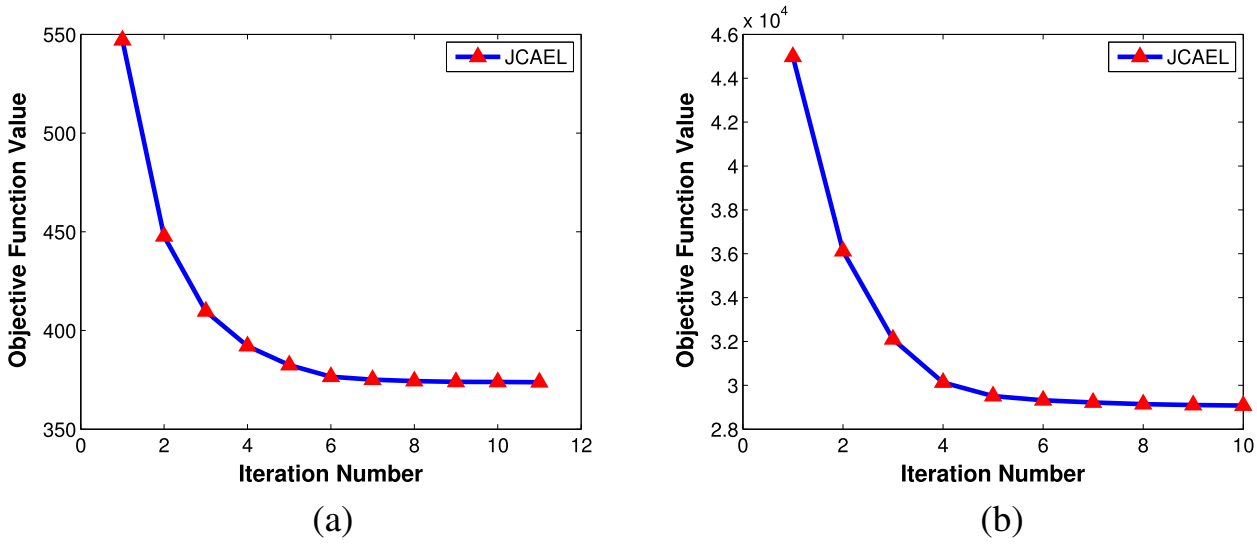

Fig. 3 Convergence curves of the objective function value in Eq. (7) using Algorithm 1. **a** HMDB13←EID. **b** UCF10←AID

HMDB13←EID, the result is worse than using all features. (2) The results arrive at the peak level when using 720 features for HMDB13←EID and using 840 features for UCF10←AID. The variance shown on the two datasets are related to the properties of the datasets. (3) After all the features are selected, the results are lower than selecting 720 features for HMDB13←EID and 840 features for UCF10←AID. In conclusion, our method reduces noises, as the results improve on both databases.

3.8 Experiment on embedding features

In this section, we would like to investigate the influence of embedding features with different dimensions. We conduct the experiment on both HMDB13←EID and UCF10←AID when the ratio of labeled video data is set to 5%. With videos being classified by LMCSVM, Fig. 6 shows the performance variation w.r.t the number of selected features. From the illustrated results, two observations can be made: (1) the result arrives at the peak level when using 390 embedding features for HMDB13←EID and 10 embedding features for UCF10←AID. The variance shown on the two datasets are seen to be related to the properties of the datasets. (2) Without embedded learning, the results is lower than using 390 embedding features for HMDB13←EID and 10 embedding features for UCF10←AID, even when all the features are used. In conclusion, our proposed JCAEL can achieve good performance due to the fact that the most valuable information

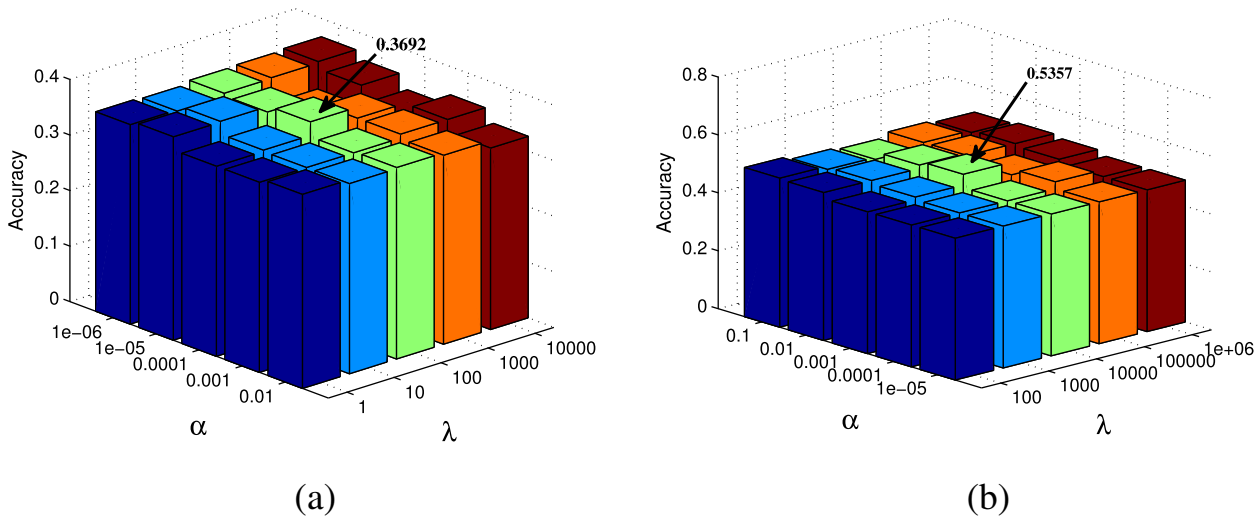

Fig. 4 The performance variance w.r.t. α and λ. **a** HMDB13←EID. **b** UCF10←AID

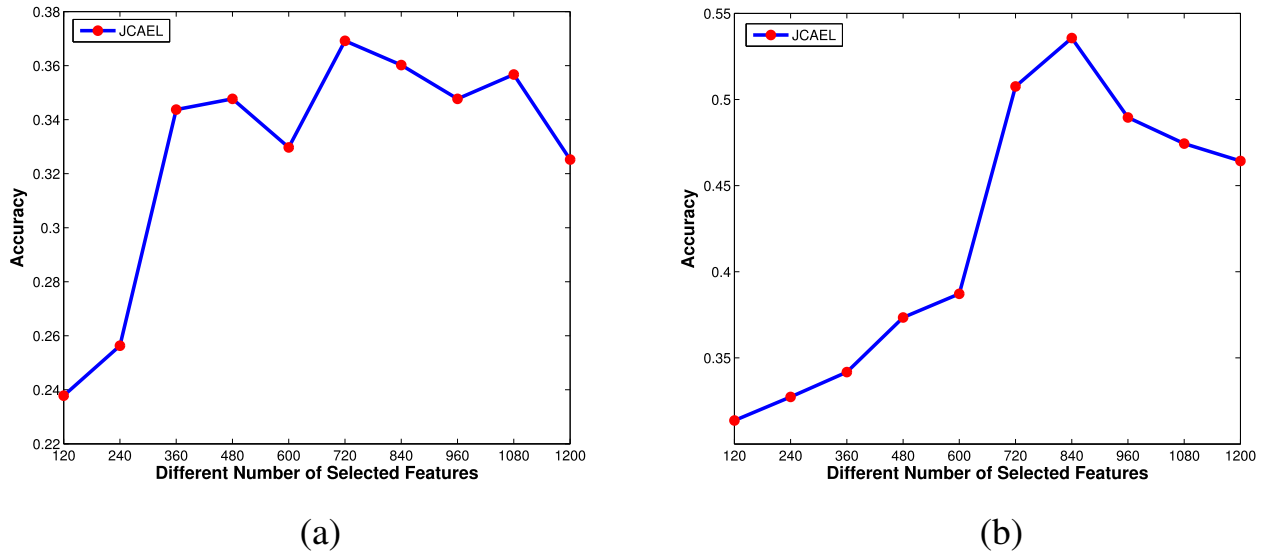

Fig. 5 The performance variation w.r.t the number of selected features using our feature selection algorithm. **a** HMDB13←EID. **b** UCF10←AID

is retained and the feature redundancies are eliminated in embedded learning.

3.9 Influence of cross-media analysis and embedded learning

To further investigate the effectiveness of the integrated cross-media analysis and embedded learning, we construct three new algorithms: (1) embedded learning part (ELP), which is the unsupervised embedded learning part of JCAEL (i.e., Eq. (6)). ELP utilizes both labeled and unlabeled videos as the training dataset, and the augmented feature is used to represent each video by ELP. (2) Cross-media analysis part (CAP), which is the transfer learning part of JCAEL (i.e., Eq. (4)). CAP transfers the knowledge from images to labeled videos, and only the still feature is used to transfer the knowledge by CAP.

We construct a new experiment to compare JCAEL with ELP and CAP on UCF←PPMI4 dataset. Other experiment setup is similar to those described in Section 3.2, and the comparison results are shown in Table 6 and Fig. 7.

From the results presented in Table 6 and Fig. 7, we can make the following observations: (1) Among different methods and different labeled ratios, JCAEL perform best. It achieves the highest accuracy in most cases, especially when only few labeled training videos

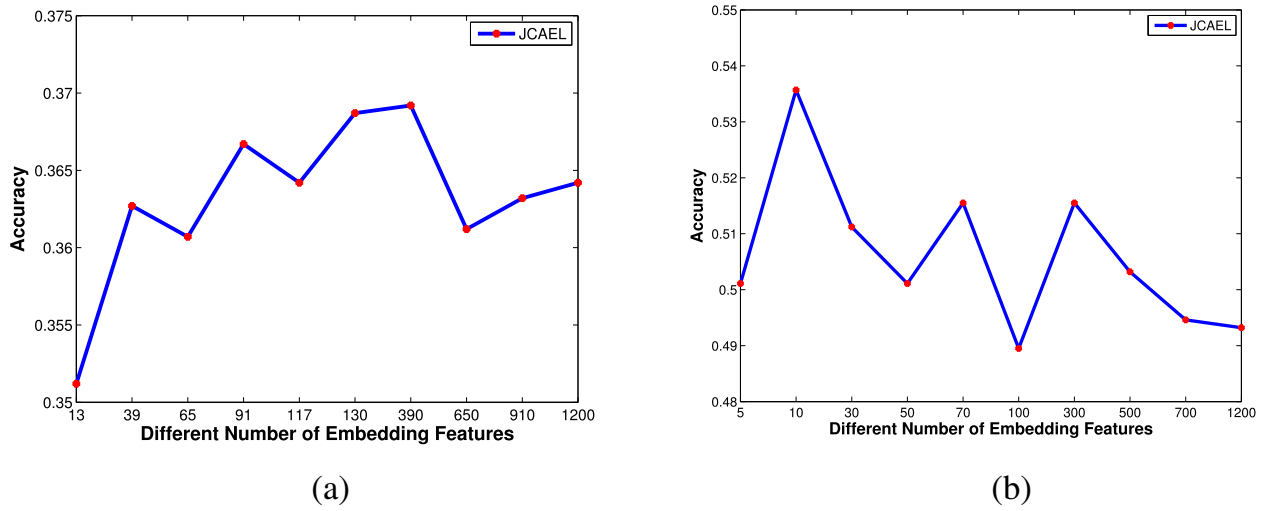

Fig. 6 The performance variation w.r.t the number of embedding features using our feature selection algorithm. **a** HMDB13←EID. **b** UCF10←AID

Table 6 Comparisons of feature selection algorithms on UCF←PPMI4 in terms of average accuracy using three classifiers when the ratio of labeled video data are set to 5%

classifiers	ELP	CAP	JCAEL
LMCSVM	***0.5537***	0.5456	**0.5804**
LSR	***0.3958***	0.3944	**0.4085**
MCkNN	***0.4085***	0.3988	**0.4433**

are provided. This is mainly due to the fact that (1) JCAEL benefits from the unsupervised embedded learning which can utilize both labeled and unlabeled data, (2) JCAEL leverages the knowledge from images to boost its performances, and (3) JCAEL integrates transfer learning and embedded learning into a joint optimization framework. In this way, gains from optimization are augmented. (2) The performance of JCAEL is generally better than that of ELP for all the labeled ratios, indicating that the JCAEL is able to use the extra knowledge from images to achieve higher accuracy. (3) JCAEL generally outperforms CAP, indicating that it is beneficial to utilize unlabeled videos for video semantic recognition, especially when the number of labeled data is not sufficient.

To show the influence of the knowledge transferred from images, we shown the confusion matrices of ELP, CAP, and JCAEL when the ratio of labeled video data are set to 5%. The confusion matrices are shown in Fig. 8. Compared with ELP and CAP, JCAEL obtains better results on "playing cello," "playing flute," "playing guitar," and "playing violin." The main reasons can be highlighted as follows: (1) the extra-related semantic knowledge is adapted from images to videos and used to obtain the coherent semantics in videos. (2) The unlabeled videos also include more relevant information, which plays positive roles in improving the performance of semantics recognition.

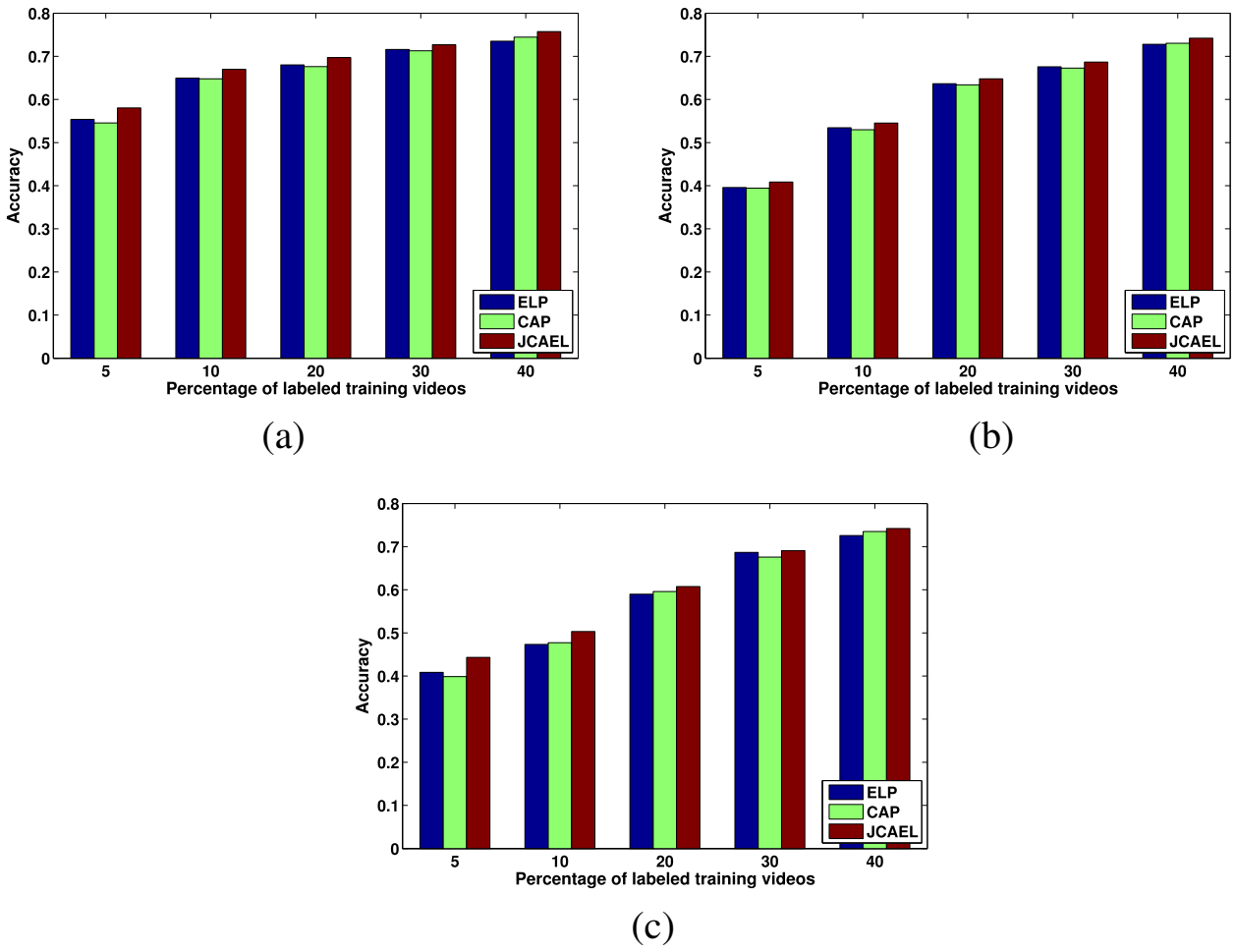

Fig. 7 Classification results of different methods on UCF←PPMI4 when the ratios of labeled video data are different. **a** LMCSVM as classifier. **b** LSR as classifier. **c** MCkNN as classifier

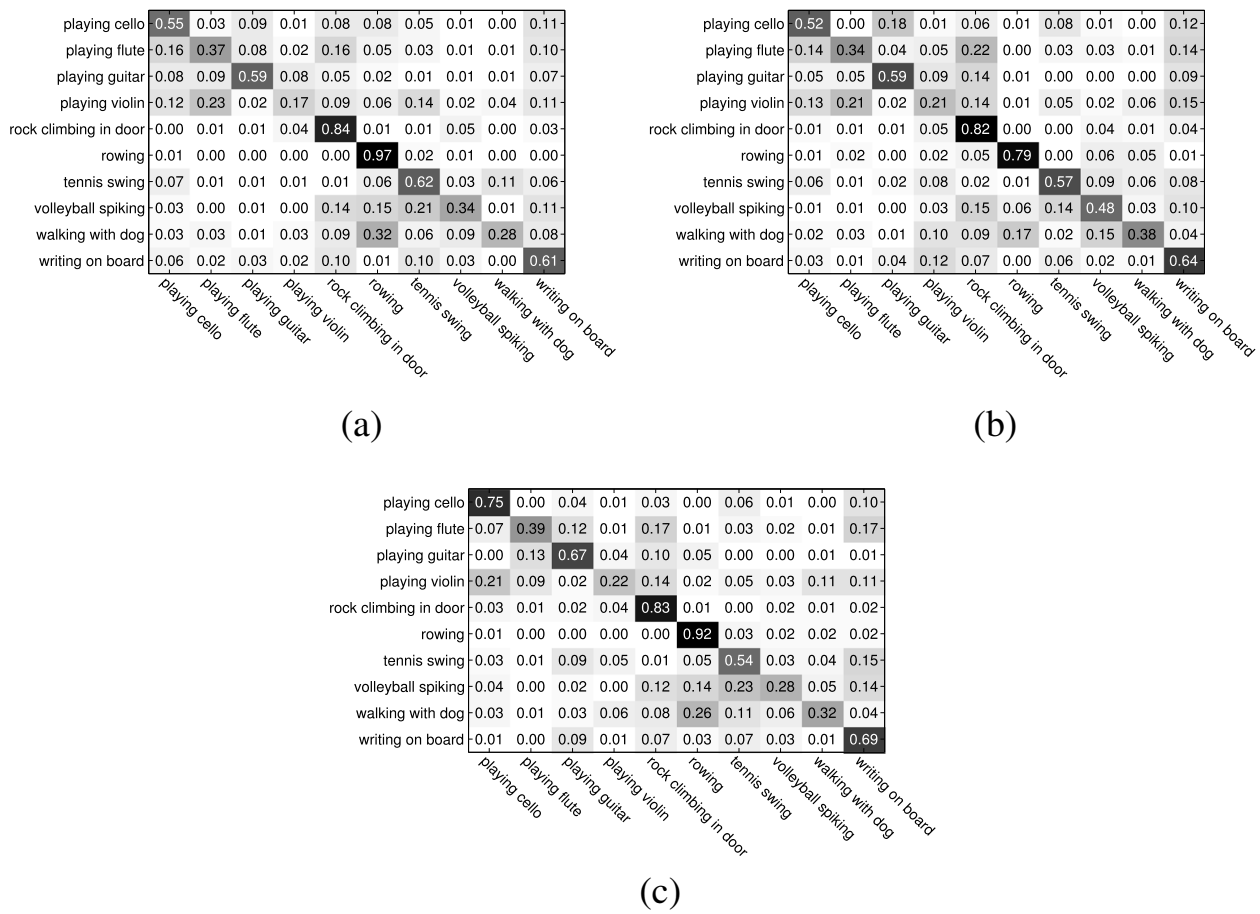

Fig. 8 Confusion matrices of ELP, CAP, and JCAEL on UCF←PPMI4 when the ratio of labeled video data are set to 5%. **a** Confusion matrices of ELP. **b** Confusion matrices of CAP. **c** Confusion matrices of JCAEL

4 Conclusions

There are many labeled images and unlabeled videos in real world. To achieve good performance for video semantic recognition, we propose a new feature selection framework, which can borrow the knowledge transferred from images to achieve its performance improvements. Meanwhile, it can utilize both labeled and unlabeled videos to enhance the performance of semantic recognition in videos. Extensive experiments validate that the knowledge transferred from images and the information contained in unlabeled videos can be used indeed to select more discriminative features, leading to the enhancement of recognition accuracies of semantics inside videos. In comparison with the existing state of the arts, the experimental results show that the proposed JCAEL has better performances in video semantics recognition. Even under the circumstance that only a few labeled training videos are available, our proposed JCAEL still performs competitive among all the compared existing state of the arts, leading to a high level of flexibility for its applications in real world.

Abbreviations

AA: Average accuracy; BoW: Bag-of-words; EnFS: Ensemble feature selection; FSFS: Fisher score feature selection; FSNM: Feature selection via joint $\ell_{2,1}$-norm minimization; JCAEL: Jointing cross-media analysis and embedded learning; JELSR: Joint embedding learning and sparse regression; LMCSVM: Linear multi-class SVM; LSR: Least square regression; LSR_{21}: $\ell_{2,1}$-norm least square regression; MCkNN: Multi-class kNN; PCA: Principal component analysis; SIFT: Scale-invariant feature transform; STIP: Space-time interest points; SVM_{21}: Multi-class $\ell_{2,1}$-norm support vector machine

Acknowledgements

Not applicable.

Funding

This work was supported in part by National Science Foundation of China (under Grant No. 61620106008, 61702165, U1509206, 61472276, 61876130). This work was supported in part by the Hebei Provincial Natural Science Foundation, China (under Grant No. F2016111005). This work was supported in part by the Foundation for Talents Program Fostering of Hebei Province (No. A201803025). This work was supported in part by Shenzhen Commission for Scientific Research & Innovations (under Grant No. JCYJ20160226191842793). This work was supported in part by Tianjin Natural Science Foundation (No. 15JCYBJC15400). This work was supported in part by the Project of Hebei Province Higher Educational Science and Technology Research (under Grant No. QN2017513). This work was supported in part by the Research Foundation for Advanced Talents of Hengshui University (under Grant No. 2018GC01).

Availability of data and materials

Please contact author for data requests.

Authors' contributions
All authors took part in the work described in this paper. The team conducted literature reading and discussion together. The author JZ designed the proposed algorithm and made the theoretical derivation of mathematics. The author ZZ collected the image and video datasets and preprocessed them, then DA, JL, and ZS together programed to implement and verify the proposed algorithm. The author JZ wrote the first version of this paper, and then, the author YH and JJ repeatedly revised the manuscript. To accomplish the final manuscript submitted, all authors participated into discussion. All authors read and approved the final manuscript.

Authors' information
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]College of Mathematics and Computer Science, Hengshui University, Hengshui, China. [2]College of Computer Science & Software Engineering, Shenzhen University, Shenzhen, China. [3]College of Intelligence and Computing, Tianjin University, Tianjin, China. [4]Hong Kong Applied Science and Technology Research Institute, Shatin, Hong Kong. [5]Office of Academic Affaires, Hengshui University, Hengshui, China. [6]Office of Academic Research, Hengshui University, Hengshui, China.

Received: 2 November 2018 Accepted: 16 January 2019
Published online: 13 February 2019

References

1. Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, X. Zhou, Semisupervised feature selection via spline regression for video semantic recognition. IEEE Trans. Neural Netw. Learn. Syst. **26**(2), 252–264 (2015)
2. L. Maddalena, A. Petrosino, Stopped object detection by learning foreground model in videos. IEEE Trans. Neural Netw. Learn. Syst. **24**(5), 723–735 (2013)
3. Y. Xu, Y. Han, R. Hong, Q. Tian, Sequential video vlad: Training the aggregation locally and temporally. IEEE Trans. Circ. Syst. Video Technol. **27**(10), 4933–4944 (2018)
4. X. Zhen, L. Shao, D. Tao, X. Li, Embedding motion and structure features for action recognition. IEEE Trans. Circ. Syst. Video Technol. **23**(7), 1182–1190 (2013)
5. S. Zhao, Y. Liu, Y. Han, R. Hong, Pooling the convolutional layers in deep convnets for action recognition. IEEE Trans. Circ. Syst. Video Technol. **28**(8), 1839–1849 (2018)
6. Y. Han, Y. Yang, F. Wu, R. Hong, Compact and discriminative descriptor inference using multi-cues. IEEE Trans. Image Process. **24**(12), 5114–5126 (2015)
7. G. Lan, C. Hou, F. Nie, T. Luo, D. Yi, Robust feature selection via simultaneous sapped norm and sparse regularizer minimization. Neurocomputing. **283**, 228–240 (2018)
8. Y. Yang, Z. Ma, A. G. Hauptmann, N. Sebe, Feature selection for multimedia analysis by sharing information among multiple tasks. IEEE Trans. Multimedia. **15**(3), 661–669 (2013)
9. Z. Ma, Y. Yang, N. Sebe, A. G. Hauptmann, Knowledge adaptation with partiallyshared features for event detectionusing few exemplars. IEEE Trans. Pattern. Anal. Mach. Intell. **36**(9), 1789–1802 (2014)
10. C. Hou, F. Nie, H. Tao, D. Yi, Multi-view unsupervised feature selection with adaptive similarity and view weight. IEEE Trans. Knowl. Data Eng. **29**(9), 1998–2011 (2017)
11. Z. Ma, F. Nie, Y. Yang, J. R. Uijlings, N. Sebe, A. G. Hauptmann, Discriminating joint feature analysis for multimedia data understanding. IEEE Trans. Multimedia. **14**(6), 1662–1672 (2012)
12. C. Deng, X. Liu, C. Li, D. Tao, Active multi-kernel domain adaptation for hyperspectral image classification. Pattern Recogn. **77**, 306–315 (2018)
13. Y. Yang, Y. Yang, H. T. Shen, Effective transfer tagging from image to video. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM). **9**(2), 14 (2013)
14. L. Duan, D. Xu, I. W.-H. Tsang, J. Luo, Visual event recognition in videos by learning from web data. IEEE Trans. Pattern. Anal. Mach. Intell. **34**(9), 1667–1680 (2012)
15. W. Zhuge, C. Hou, F. Nie, D. Yi, in *Pattern Recognition (ICPR), 2016 23rd International Conference On*. Unsupervised feature extraction using a learned graph with clustering structure (IEEE, Cancun, 2016), pp. 3597–3602
16. J. Zhang, Y. Han, J. Tang, Q. Hu, J. Jiang, in *Proceedings of the 22nd ACM International Conference on Multimedia*. What can we learn about motion videos from still images? (ACM, Orlando, 2014), pp. 973–976
17. J. Zhang, Y. Han, J. Jiang, Tensor rank selection for multimedia analysis. J. Vis. Commun. Image Represent. **30**, 376–392 (2015)
18. Y. Han, J. Zhang, Z. Xu, S.-I. Yu, in *Proceedings of the 17th AAAI Conference on Late-Breaking Developments in the Field of Artificial Intelligence*. Discriminative multi-task feature selection (AAAI Press, Bellevue, 2013), pp. 41–43
19. S. Ji, L. Tang, S. Yu, J. Ye, A shared-subspace learning framework for multi-label classification. ACM Trans. Knowl. Disc. Data. **4**(2), 1–29 (2010)
20. X. Cai, F. Nie, H. Huang, C. Ding, in *IEEE International Conference on Data Mining*. Multi-class l2,1-norm support vector machine (IEEE, Brussels, 2012), pp. 91–100
21. J. Zhang, Y. Han, J. Tang, Q. Hu, J. Jiang, Semi-supervised image-to-video adaptation for video action recognition. IEEE Trans. Cybern. **47**(4), 960–973 (2016)
22. F. Nie, H. Huang, X. Cai, C. Ding, in *International Conference on Neural Information Processing Systems*. Efficient and robust feature selection via joint l2,1-norms minimization (Curran Associates Inc, Vancouver, 2010), pp. 1813–1821
23. X. He, D. Cai, P. Niyogi, in *Advances in Neural Information Processing Systems*. Laplacian score for feature selection (Curran Associates Inc, Vancouver, 2006), pp. 507–514
24. J. Zhang, J. Jiang, Rank-optimized logistic matrix regression toward improved matrix data classification. Neural Comput. **30**(2), 1–21 (2018)
25. Y. Han, Y. Yang, X. Zhou, in *IJCAI, vol. 13*. Co-regularized ensemble for feature selection, (Beijing, China, 2013), pp. 1380–1386
26. C. Hou, F. Nie, D. Yi, Y. Wu, in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence, vol. 22*. Feature selection via joint embedding learning and sparse regression, (Barcelona, 2011), pp. 1324–1329
27. H. Kuehne, H. Jhuang, R. Stiefelhagen, T. Serre, in *IEEE International Conference on Computer Vision, ICCV 2011, November*. Hmdb51: A large video database for human motion recognition (IEEE, Barcelona, 2012), pp. 2556–2563
28. B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, F. F. Li, in *International Conference on Computer Vision*. Human action recognition by learning bases of action attributes and parts (IEEE, Barcelona, 2011), pp. 1331–1338
29. N. Ikizler, R. G. Cinbis, S. Pehlivan, P. Duygulu, in *International Conference on Pattern Recognition*. Recognizing actions from still images (IEEE, Tampa, 2008), pp. 1–4
30. O. Deniz, I. Serrano, G. Bueno, T. K. Kim, in *International Conference on Computer Vision Theory and Applications*. Fast violence detection in video (IEEE, 2015), pp. 478–485
31. A. Gupta, A. Kembhavi, L. S. Davis, Observing human-object interactions: using spatial and functional compatibility for recognition. IEEE Trans. Pattern. Anal. Mach. Intell. **31**(10), 1775–89 (2009)
32. B. Yao, F. F. Li, in *Computer Vision and Pattern Recognition*. Grouplet: A structured image representation for recognizing human and object interactions (IEEE, San Francisco, 2010), pp. 9–16
33. V. Delaitre, I. Laptev, J. Sivic, in *British Machine Vision Conference, BMVC 2010, August 31 - September 3, 2010. Proceedings*. Recognizing human actions in still images: a study of bag-of-features and part-based representations, (Aberystwyth, 2010), pp. 1–11
34. X. Liu, Z. Li, C. Deng, D. Tao, Distributed adaptive binary quantization for fast nearest neighbor search. IEEE Trans. Image Process. **26**(11), 5324–5336 (2017)
35. X. Liu, J. He, B. Lang, Multiple feature kernel hashing for large-scale visual search. Pattern Recog. **47**(2), 748–757 (2014)
36. X. Liu, J. He, S. F. Chang, Hash bit selection for nearest neighbor search. IEEE Trans. Image Process. **26**(11), 5367–5380 (2017)
37. R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification (2nd Edition)*. (Wiley, 2001), pp. 55–88