Semisupervised Discriminant Multimanifold Analysis for Action Recognition

Zengmin Xu, Ruimin Hu, *Senior Member, IEEE*, Jun Chen, Chen Chen, Junjun Jiang, Jiaofen Li, and Hongyang Li

***Abstract*— Although recent semisupervised approaches have proven their effectiveness when there are limited training data, they assume that the samples from different actions lie on a single data manifold in the feature space and try to uncover a common subspace for all samples. However, this assumption ignores the intraclass compactness and the interclass separability simultaneously. We believe that human actions should occupy multimanifold subspace and, therefore, model the samples of the same action as the same manifold and those of different actions as different manifolds. In order to obtain the optimum subspace projection matrix, the current approaches may be mathematically imprecise owe to the badly scaled matrix and improper convergence. To address these issues in unconstrained convex optimization, we introduce a nontrivial spectral projected gradient method and Karush–Kuhn–Tucker conditions without matrix inversion. Through maximizing the separability between different classes by using labeled data points and estimating the intrinsic geometric structure of the data distributions by exploring unlabeled data points, the proposed algorithm can learn global and local consistency and boost the recognition performance. Extensive experiments conducted on the realistic video data sets, including JHMDB, HMDB51, UCF50, and UCF101, have demonstrated that our algorithm outperforms the compared algorithms, including deep learning approach when there are only a few labeled samples.**

***Index Terms*— Discriminant analysis, Karush–Kuhn–Tucker (KKT) conditions, manifold learning, semisupervised learning, spectral projected gradient (SPG).**

Manuscript received July 18, 2017; revised March 8, 2018 and August 27, 2018; accepted November 27, 2018. Date of publication February 13, 2019; date of current version September 18, 2019. This work was supported in part by the National Nature Science Foundation of China under Grant U1736206, Grant U1611461, Grant 61862015, Grant 61501413, Grant 11761024, Grant 11561015, Grant 61231015, Grant 61367002, Grant 61502152, and Grant 61671336, in part by the Technology Research Program of Ministry of Public Security under Grant 2016JSYJA12, in part by the Hubei Province Technological Innovation Major Project under Grant 2017AAA123 and Grant 2016AAA015, in part by the National Key Research and Development Program of China under Grant 2017YFC0803700, in part by the Guangxi Key Research and Development Program under Grant AB17195025, in part by the Nature Science Foundation of Guangxi Province under Grant 2016GXNSFFA380009 and Grant 2016GXNSFAA380074, and in part by the Guangxi Young Teachers' Basic Ability Improvement Project under Grant 2017KY0190. *(Corresponding author: Ruimin Hu.)*

Z. Xu is with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China, also with the Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China, also with the Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China, and also with the School of Mathematics and Computing Science, Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation, Guilin University of Electronic Technology, Guilin 541004, China (e-mail: zengminxu@gmail.com).

R. Hu, J. Chen, and H. Li are with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China, also with the Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China, and also with the Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China (e-mail: hurm1964@gmail.com; chenj@whu.edu.cn; lihy@whu.edu.cn).

C. Chen is with the Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, Charlotte, NC 28223 USA (e-mail: chenchen870713@gmail.com).

J. Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China, and also with the Peng Cheng Laboratory, Shenzhen, China (e-mail: junjun0595@163.com).

J. Li is with the School of Mathematics and Computing Science, Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation, Guilin University of Electronic Technology, Guilin 541004, China (e-mail: zengminxu@gmail.com; lixiaogui1290@163.com).

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

Digital Object Identifier 10.1109/TNNLS.2018.2886008

I. Introduction

VISUAL recognition draws strong research interest in computer vision because of its promising applications for feature selection, image annotation, video concept detection, and so on [1]–[29]. With the developments in cloud storage technologies, the number of personal images/videos increases rapidly, and it becomes an important challenge to organize these resources effectively. Common approaches of visual recognition are to train supervised classifiers from large-scale labeled data. However, the amount of labeled data is extremely scarce compared with the unlabeled data in the real world. When confronted with huge amounts of unlabeled samples, manual annotation or labeling should be prohibitive. Consequently, semisupervised learning, which can make good use of both labeled and unlabeled data, is applied to explore feature correlation from the original feature space.

Motivated by the progress of semisupervised learning, a few research attention has been paid to semisupervised action recognition [30], [31]. A common limitation of the existing supervised and semisupervised action recognition algorithms is that they evaluate the importance of commonly shared structure between different actions, without considering intraclass compactness and interclass separability simultaneously [30], [31]. For example, even though legs motion appears in similar actions such as the SoccerJuggling and the SoccerPenalty, these between-class actions have much similar motion and dissimilar components simultaneously. Although the shared structural uncovering and label correlation mining have proven beneficial to action recognition in [30] and [31], the ways to learn discriminant features in a semisupervised framework for action recognition have not been largely addressed. To solve this problem, some state-of-the-art

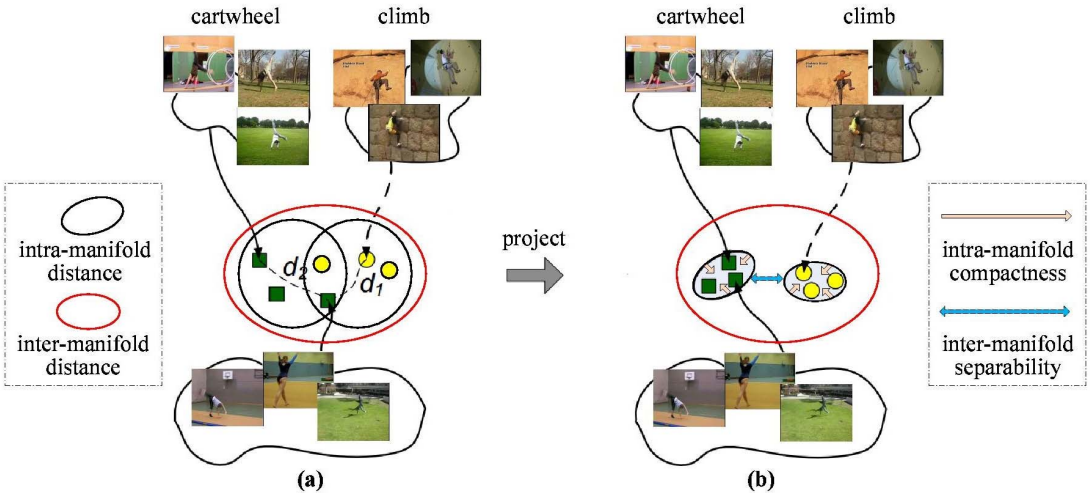


Fig. 1. Illustration of the proposed approach. (a) In feature space, actions can be described as points. However, intersection formed by different classes may confuse the discriminative. (b) By employing discriminant multimanifold analysis, points in feature space can be mapped into new manifolds where different actions are well separated while the same actions become closer. The proposed method not only preserves local geometrical properties but also maximize the discriminatory power between classes on multimanifold by exploiting within-class and between-class similarity graphs.

algorithms are proposed to take a discriminant analysis into consideration for visual recognition. For example, the works in [32]–[36] implement their methods in a supervised way.

Another limitation of current semisupervised approaches is that they solve their nonconvex optimization by impressive derivation and alternating-least-squares-like iterative algorithm, which fails to discover the most valuable optimum in a mathematical way [3], [4], [30], [31], [37]. This is because the subproblem of objective function optimization is less rigorous, which have not discussed the singularity of the deduced matrix. In addition, the optimum is supposed to satisfy the Karush–Kuhn–Tucker (KKT) conditions, but they do not explain the KKT conditions of orthogonal constraint, and the accuracy of convergence optimum also lack of further analysis. Recent research studies have indicated that it is beneficial to obtain an optimal solution by projected gradient methods. Motivated by this fact, the projected gradient method has been introduced to the field of multimedia [38], [39]. Although spectral projected gradient (SPG) method [40] has been studied extensively in both theory and practice [38], [41], [42], so far no study has formally applied its techniques to action recognition in semisupervised way.

As mentioned above, it remains unclear how to manually define feature correlation in action recognition. Thus, we propose to model the intramanifold compactness and the intermanifold separability simultaneously and characterize high-level semantic pattern through the local action features by discriminant multimanifold analysis, as shown in Fig. 1. The proposed algorithm combines the strengths of semisupervised learning, discriminant analysis, multitask learning, and unconstrained optimization. Both labeled and unlabeled data are utilized for action recognition in classifiers' training phase.

A. Motivation and Contributions

It is true that there is a trend to apply deep learning approaches to achieve good action recognition performance, by relying on large-scale labeled training data. Although there are a few large-scale data sets, e.g., Sports1M [43], YouTube8M [44], and ActivityNet [45], obtaining and annotating such data sets require a significant amount of time, resources, and effort. In contrast, collecting unlabeled videos is much easier. The semisupervised learning can effectively leverage the unlabeled data.

Moreover, videos in those data sets are limited to sports and/or daily activities. For real-world applications such as anomaly detection in surveillance and labeled data (videos contain rare anomalous events, e.g., crime related activities) are notoriously hard to obtain [46].

In addition, videos in surveillance applications are very different from other web-based multimedia videos, e.g., Sports1M, YouTube8M, and ActivityNet, due to content, background, device noise, action complexity, viewpoint, scale, and so on (see Fig. 2). The deep learning model on multimedia data set may not work well on surveillance data set, as the deep learning approach learned to exploit the specifics of a particular action from multimedia videos rather than the learning models of characters that are then used for parsing the action from other types of videos [47].

The deep learning approaches not only rely on large-scale labeled training data but also are restricted by the capacity of GPUs. Moreover, overfitting is still an unsolved problem, especially when there is limited training data. How to leverage the unlabeled data and how to pursue efficient learning methods trained on small data sets are worthwhile scientific questions of broad interest to the community [19], [48], [49].

The goal of this paper is to uncover the discriminative information by exploring action features and achieve the state-of-the-art action recognition performance based on the semi-supervised setting, which uses only part of the labeled training data, as compared with other semisupervised approaches under the same setting. The distinction we want to make is that we do not aim to compete with fully supervised approaches,

which reported better performance than ours on the evaluated data sets. The contributions are summarized as follows:

1) Ours is the first work to consider both multimanifold analysis and semisupervised learning in action recognition, given that samples (action videos) may lie in a multimanifold subspace. By modeling a multimanifold subspace, both intraclass compactness and interclass separability are taken into account.
2) To solve the unconstrained convex optimization in our problem, we propose to incorporate SPG and KKT conditions to avoid matrix inversion, as done in [26] and [27], which may suffer from the singularity, thereby leading to better convergence and a more accurate solution. In addition, we provide experimental justification on the convergence.
3) We not only introduce a new idea (i.e., multimanifold analysis) in the problem formulation of semisupervised action recognition but also develop an effective and efficient algorithm to solve the optimization of the objective function.
4) Extensive experiments have validated that our method achieves the best recognition performances on four benchmarks in the semisupervised setting, while has the fastest training speed as compared with the state of the art [e.g., subfeature uncovering with sparsity (SFUS), semisupervised feature correlation mining (SFCM), and multiple feature correlation uncovering (MFCU)]. We believe our work provides valuable insights into video action analysis in a semisupervised manner.

II. Related Work

In this section, we review the related research on manifold learning, semisupervised learning, and multitask learning.

A. Discriminant Analysis

Previous works have stated that manifold learning is capable of mining geometry structures information by regarding a space of probabilities as a manifold [33], [34], [36], [50]–[55].

Cai and He [33] perform an active learning algorithm which lies on the data manifold adaptive kernel space by using graph Laplacian, which can reflect the underlying geometry of the data. Harandi *et al.* [51] develop a discriminant analysis approach on Grassmannian manifolds by characterizing intraclass compactness and interclass separability. Li *et al.* [52] contribute a novel coclustering algorithm based on symmetric nonnegative matrix trifactorization by manifold ensemble learning. Yan *et al.* [53] propose a novel multitask learning framework for multiview action recognition by multitask linear discriminant analysis. Jiang *et al.* [34]–[36] try to match a low resolution or poor quality face image to a gallery of high-resolution face images by discriminant analysis on multimanifold. Yu and Zhao [50], [54] introduce the penalty of a lasso or elastic net into the exponential discriminant analysis so that the key variables responsible for fault diagnosis can be automatically selected. In [56]–[58], they exploit the local manifold structure to capture the discrimination features when reconstructing the face images. Ma *et al.* [55] exploit the intrinsic geometrical structures among the feature points for shape registration based on manifold regularization. Inspired by these research studies, we try to join the idea of discriminant analysis into a semisupervised framework and use the labeled data points to maximize the separability between different classes.

B. Semisupervised Learning

Semisupervised learning has been widely used for its promising performance in different applications [1], [3], [4], [30], [31], [37], [59]–[62]. Given labeling a large amount of training data is time-consuming and expensive, unlabeled samples can be exploited to learn data correlation by semisupervised learning. Thus, semisupervised learning is beneficial in terms of both the data analysis performance and human laboring cost.

Graph Laplacian-based semisupervised learning has shown its simplicity and efficiency in visual concept recognition [63]. Nie *et al.* [59] propose a manifold learning framework based on graph Laplacian and compared its performance with other algorithms. Ma *et al.* [4] develop a novel feature selection method and apply it to automatic image annotation. Yang *et al.* [1] present a framework for multimedia content analysis and retrieval which consists of two independent algorithms. Chang and Yang [3] build a semisupervised feature selection framework by mining correlations among multiple tasks and apply it to different multimedia applications. Wang *et al.* [30], [31] point out that action recognition can be improved by a complicated formulation and iterative algorithm. In addition, semisupervised learning has also been applied to solving the problems of face recognition [60], image matching [61], image fusion [62], and so on. Motivated by these papers, we design a semisupervised learning algorithm with graph embedding discriminant analysis, then the intrinsic geometric structure of the data distributions can be estimated by exploring the unlabeled data points.

C. Multitask Learning

Multitask learning has gained increasing interest in many applications for its advantage, which can learn multiple related tasks with a shared representation [2], [64], [65]. Recent research studies have indicated that learning multiple related tasks jointly always outperforms learning them independently. Inspired by the progress of multitask learning, researchers have introduced it to the field of multimedia and demonstrated its promising performance on multimedia analysis. For example, Yang *et al.* [2] study a novel multitask feature selection algorithm in a batch mode by leveraging shared information among multiple related tasks. Ma *et al.* [64] design a multitask learning framework to jointly optimize the classifiers for both laboratory and real-world data sets. Yang *et al.* [65] learn a novel clustering model to capture correlations among the related clustering tasks and/or within an individual task. Despite their good performances, these classical algorithms are all implemented only with labeled training data. Following the related works, the proposed framework can learn the

Fig. 2. Example frames from (a) Current multimedia videos. (b) TRECVID surveillance videos. (c) City surveillance videos. We can see the characteristics in different types of videos. The left two columns show the widely used multimedia videos from movies, sports, youtube, and so on. The middle two columns illustrate multicamera airport surveillance domain evaluation data from TRECVID. The right two columns demonstrate various surveillance camera records from city environment consisting of indoor and outdoor scenes. Given the realistic application of video investigation, finding specific abnormal actions in daily life should be focused on public safety. Nevertheless, existing researches including action recognition and event detection have not contained those rare unusual actions.

global consistency and the local geometric, and hence, the performance can be improved by mining correlations between multiple related tasks.

There are two aforementioned research studies close to our work, which are both proposed by Wang *et al.* [30], [31]. They assume that the samples from different actions define a single data manifold in the feature space, visual words of different action videos may share a common structure in a low-dimensional space. They introduce a transformation matrix Q to characterize the shared information and employ a regularize term of the shared information among different features. They solve their constrained nonconvex optimization problem by comprehensive derivation and alternating-least-squares-like iterative algorithm. However, the deduced inverse matrix is close to singular or badly scaled during the optimization process, which may make the results inaccurate.

To solve the above-mentioned issues, we model the samples of the same action as the same manifold and those of different actions as different manifolds. As described before, we claim that multimanifold mapping can maximize the discriminatory power while preserving local geometry, mining shared structure is not our purpose, so we discard the shared structure regularization term, and model the local geometrical structure of manifolds by building a within-class similarity graph A_w and a between-class similarity graph A_b. We also remove the selection matrix U in our function. Since the proposed optimization solution in [30] and [31] may be mathematically imprecise, we introduce the SPG method and the KKT conditions to avoid matrix inversion and improper convergence.

III. Proposed Approach

This section begins with an elaboration of the formulation of the proposed approach. Our method incorporates several techniques including the least-square loss function, graph-based semisupervised learning, feature correlation mining, SPG, KKT, and discriminant multimanifold analysis. It is named semisupervised discriminant multimanifold analysis (SDMM). Following this, we describe how to obtain the classifiers in detail.

A. Formulation

To exploit the feature correlation for action recognition, we define the training set as $X = [x_1, \ldots, x_n] \in \mathbb{R}^{d\times n}$ and then associate it with its ground truth labels matrix $Y = [y_1, \ldots, y_n]^T \in \{0, 1\}^{n\times c}$. Note that $x_i \in \mathbb{R}^{d\times 1}$ is the i-th datum, and n is the size of X. We aim to learn c prediction functions (classifiers) $\{f_\ell\}_{\ell=1}^{c}$, with one for each class. c stands for the class number. Usually, the prediction function f is defined as

$$f(x) = w^T x \tag{1}$$

where x is a datum and $w \in \mathbb{R}^{d\times 1}$ is weight vectors. By denoting $W = [w_1, \ldots, w_c] \in \mathbb{R}^{d\times c}$, the above-mentioned function becomes

$$f(X) = X^T W. \tag{2}$$

As indicated in [66], the least-square loss function achieves comparable performance to other loss functions, e.g., hinge loss or logistic loss. To obtain the projection matrix W, we employ least square regression to solve the following optimization problem:

$$\min_W \|X^T W - Y\|_F^2 + \alpha \|W\|_F^2 \tag{3}$$

where α is the regularization parameters. $\|\cdot\|_F^2$ denotes Frobenius norm. $\|W\|_F^2$ controls the complexity of the model to avoid overfitting.

Following the assumption of [31], the nearby data points are likely to have the same label, and the edges of graph A refers to connect pairs of data points. A denotes the symmetric matrix with elements describing the similarity between the pairs of data points. However, unlike [31], utilizing one graph model to approximate the density and manifold information,

in this paper, we model the local geometrical structure of manifolds by building a within-class similarity graph A_w and a between-class similarity graph A_b. For simplicity, A_w and A_b are defined based on the nearest neighbor graphs as follows:

$$A_w(i,j) = \begin{cases} 1, & x_i \in N_w(x_j) \text{ or } x_j \in N_w(x_i) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$A_b(i,j) = \begin{cases} 1, & x_i \in N_b(x_j) \text{ or } x_j \in N_b(x_i) \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

In (4), $N_w(x_j)$ is the set of neighbors x_j, sharing the same label with x_i. $N_b(x_i)$ contains some neighbors having different labels in (5). We note that intraclass and interclass distances between points can be encoded on manifold by using similarity graphs [36].

B. Discriminant Analysis

Our goal is to maximize discriminatory power while preserving local geometry, by mapping the points to new manifold, i.e., $w : X_i \rightarrow F_i$. To better demonstrate the relationship between the data distribution on manifold and feature correlation mining, we define a predicted label matrix $F = [F_1, \ldots, F_n]^T \in \mathbb{R}^{n\times c}$ for all the training videos in X, where $F_i \in \mathbb{R}^{c\times 1}$ is the predicted label vector of the i-th datum $x_i \in X$.

Inspired by the manifold discriminant analysis [34], [51], [67], we aim to minimize the intramanifold compactness and maximize the intermanifold separability simultaneously. A suitable transform would place the connected points of A_w as close as possible, while moving the connected points of A_b as far as possible. This goal can be achieved by optimizing the following two objective functions:

$$f_1 = \min \frac{1}{2}\sum_{\ell=1}^{c}\sum_{i,j=1}^{n}(F_{i\ell} - F_{j\ell})^2 A_w(i,j) \tag{6}$$

$$f_2 = \max \frac{1}{2}\sum_{\ell=1}^{c}\sum_{i,j=1}^{n}(F_{i\ell} - F_{j\ell})^2 A_b(i,j) \tag{7}$$

where $F_{i\ell}$ is the ℓth element of F_i. f_1 punishes neighbors in the same class if they are mapped far away, while f_2 punishes samples of different classes if they are mapped close together. Hence, the overall discriminative information can be represented as

$$f = \frac{1}{2}\sum_{\ell=1}^{c}\sum_{i,j=1}^{n}(F_{i\ell} - F_{j\ell})^2 A_w(i,j) - \frac{1}{2}\beta\sum_{\ell=1}^{c}\sum_{i,j=1}^{n}(F_{i\ell} - F_{j\ell})^2 A_b(i,j) \tag{8}$$

where β is a regularization parameter which controls the tradeoff between the intramanifold compactness term and the intermanifold separability term. Note that

$$\begin{aligned} &\frac{1}{2}\sum_{\ell=1}^{c}\sum_{i,j=1}^{n}(F_{i\ell} - F_{j\ell})^2 A_w(i,j) \\ &\quad = \frac{1}{2}\sum_{i,j=1}^{n} A_w(i,j)\left(F_i^T F_i + F_j^T F_j - 2F_i^T F_j\right) \\ &\quad = \operatorname{tr}(F^T(D_w - A_w)F) = \operatorname{tr}(F^T L_w F) \end{aligned} \tag{9}$$

where $\operatorname{tr}(\cdot)$ denotes trace operator, D_w is a diagonal matrix with $D_w(i,i) = \sum_{j=1}^{n} A_w(i,j)$, and $L_w = D_w - A_w$ is the Laplacian matrix [68]. Similarly, (7) can be simplified to

$$\begin{aligned} &\frac{1}{2}\sum_{\ell=1}^{c}\sum_{i,j=1}^{n}(F_{i\ell} - F_{j\ell})^2 A_b(i,j) \\ &\qquad = \operatorname{tr}(F^T(D_b - A_b)F) = \operatorname{tr}(F^T L_b F) \end{aligned} \tag{10}$$

where D_b is a diagonal matrix with $D_b(i,i) = \sum_{j=1}^{n} A_b(i,j)$. Therefore, equation (8) can be rewritten as

$$\begin{aligned} f &= \frac{1}{2}\sum_{\ell=1}^{c}\sum_{i,j=1}^{n}(F_{i\ell} - F_{j\ell})^2 A_w(i,j) \\ &\quad -\frac{1}{2}\beta\sum_{\ell=1}^{c}\sum_{i,j=1}^{n}(F_{i\ell} - F_{j\ell})^2 A_b(i,j) \\ &= \operatorname{tr}(F^T(L_w - \beta L_b)F). \end{aligned} \tag{11}$$

C. Multitask Learning

To alleviate the tedious work in supervised learning, we extend the above-mentioned function to a graph-based semisupervised method for leveraging both labeled and unlabeled data as shown in [4] and [37]. Most existing semisupervised learning methods assume that the nearby data points are likely to have the same label. Specifically, the data points which can be connected via a path through high-density regions on the data manifold are likely to have the same label [4], [30], [37]. Nevertheless, the density and manifold information are inadequate due to limited labeled data. To relieve this problem, we utilize the graph model mentioned in Section III-B to approximate the density and manifold information.

To begin with, we redefine the training data set as $X = [X_l^T, X_u^T]^T$, where $X_l = [x_1, \ldots, x_m]^T$ and $X_u = [x_{m+1}, \ldots, x_n]^T$ are the two subsets of the data with labels and without labels, respectively. The label matrix of X is $Y = [Y_l^T, Y_u^T]^T$, where $Y_l = [y_1, \ldots, y_m]^T \in 0, 1^{m\times c}$ and $Y_u = [y_{m+1}, \ldots, y_n]^T \in \mathbb{R}^{(n-m)\times c}$ is a matrix with all zeros. According to [30], [37], and [69], the graph embedded label prediction matrix F should be consistent with similarity graphs A_w and A_b, and the ground-truth labels Y. The idea of multimanifold and label consistency can be generalized as

$$\begin{aligned} &\min_F \sum_{\ell=1}^{c}\left[\frac{1}{2}\sum_{i,j=1}^{n}(F_{i\ell} - F_{j\ell})^2 A_w(i,j) \right. \\ &\qquad \left. -\frac{1}{2}\beta\sum_{i,j=1}^{n}(F_{i\ell} - F_{j\ell})^2 A_b(i,j) + \sum_{i=1}^{n}(F_{i\ell} - y_{i\ell})^2\right] \\ &\Rightarrow \min_F \operatorname{tr}(F^T(L_w - \beta L_b)F) + \operatorname{tr}(F - Y)^T(F - Y). \end{aligned} \tag{12}$$

Different from previous shared structure learning algorithms [4], [30], [31], [37], we do not take shared structure learning into account in semisupervised learning framework. Instead, we propose a novel joint framework by incorporating graph embedding method on multimanifold and the classifiers, which can be formulated as

$$\min_{F,w} \operatorname{tr}(F^T(L_w - \beta L_b)F) + \operatorname{tr}(F-Y)^T(F-Y) + \mu \sum_{\ell=1}^{c}\left(\sum_{i=1}^{n} \operatorname{loss}(f_\ell(x_i), F_{i\ell}) + \alpha\|w_\ell\|^2\right) \quad (13)$$

where $\mu > 0, \alpha > 0$, and $\beta > 0$ are the regularization parameters. As discussed in Section III-A, we employ the Frobenius norm regularized loss function and then rewrite our objective as

$$\min_{F,W} \operatorname{tr}(F^T(L_w - \beta L_b)F) + \operatorname{tr}(F-Y)^T(F-Y) + \mu\left(\|X^TW - F\|_F^2 + \alpha\|W\|_F^2\right). \quad (14)$$

There are three issues worthy of consideration. First, our objective function (14) is an unconstrained convex optimization problem. It does not contain the shared subspace information regularization term, hence the global optimum can be obtained by performing alternating least squares or SPG method [38], [42]. Second, the solution of objective functions shown in [1], [4], [30], and [31] have not discussed the singularity of the matrix. In [38] and [42], the SPG method has been proved that it can handle the aforementioned issues without matrix inversion. Third, the convergence conditions in [4], [30], and [31] merely depend on monotone decreasing. Since the objective function value becoming stable may be mathematically improper convergence, we utilize KKT conditions to deal with this matter.

D. Optimization

For reducing the dimension of X, we follow [37] to perform singular value decomposition (SVD), in which all the eigenvectors corresponding to the nonzero eigenvalues of the covariance matrix is preserved.

After that, according to [38] and [42], a general unconstrained minimization problem can be solved iteratively by introducing the SPG method and the trace operator. Therefore, we define a function $g(F, W)$ as a new objective problem instead of (14)

$$g(F,W) = \min_{F,W} \operatorname{tr}(F^T(L_w - \beta L_b)F) + \operatorname{tr}(F-Y)^T(F-Y) + \mu\operatorname{tr}(X^TW - F)^T(X^TW - F) + \mu\alpha\operatorname{tr}(W^TW). \quad (15)$$

By setting the derivative of (15) with respect to F and W, respectively, we have

$$\nabla g_F = \frac{\partial g(F,W)}{\partial F} = 2(L_w - \beta L_b)F + 2(F-Y) - 2\mu(X^TW - F) \quad (16)$$

$$\nabla g_W = \frac{\partial g(F,W)}{\partial W} = 2\mu X(X^TW - F) + 2\mu\alpha W. \quad (17)$$

If (F^*, W^*) is an approximate stationary point of (15), it is supposed to meet the KKT condition of (15) like this

$$\nabla g_F\Big|_{(F^*,W^*)} = 0, \quad \nabla g_W\Big|_{(F^*,W^*)} = 0. \quad (18)$$

Then, the iteration stopping criterion becomes

$$\|\nabla g_F(F^*, W^*)\|^2 + \|\nabla g_W(F^*, W^*)\|^2 \leqslant \varepsilon \quad (19)$$

where ε is a nonnegative small constant. In summary, the classifiers training process of the proposed method is detailed in Algorithm 1.

Algorithm 1 SDMM Algorithm

Input:
The training data $X \in \mathbb{R}^{d\times n}$
The training data labels $Y \in \mathbb{R}^{n\times c}$
Semi-supervised Parameters α, β and μ.
SPG Parameters $M, \alpha^+_{\min}, \alpha^+_{\max}, \gamma, \delta_1$ and δ_2

Output:
Optimized $W^* \in \mathbb{R}^{d\times c}$

Perform SVD to reduce X's dimension according to [37]
Compute the within-class similarity graph $L_w \in \mathbb{R}^{n\times n}$
Compute the between-class similarity graph $L_b \in \mathbb{R}^{n\times n}$
Initialize $t = 0$, $\alpha_0^+ \in (\alpha^+_{\min}, \alpha^+_{\max})$, $\lambda = 1$
Initialize $F^{(0)} \in \mathbb{R}^{n\times c}$ randomly
Initialize $W^{(0)} \in \mathbb{R}^{d\times c}$ randomly

repeat ▷ SPG method
  Compute $dF^{(t)} = -\alpha_t^+ \nabla g_F(F^{(t)}, W^{(t)})$
  Compute $dW^{(t)} = -\alpha_t^+ \nabla g_W(F^{(t)}, W^{(t)})$
  Compute $\check{F} = F^{(t)} + \lambda dF^{(t)}$
  Compute $\check{W} = W^{(t)} + \lambda dW^{(t)}$
  if $g(\check{F}, \check{W}) \leqslant \gamma\lambda\{\langle dF^{(t)}, \nabla g_F(F^{(t)}, W^{(t)})\rangle + \langle dW^{(t)}, \nabla g_W(F^{(t)}, W^{(t)})\rangle\} + \max_{0\leqslant j\leqslant \min\{t, M-1\}} g(F^{(t-j)}, W^{(t-j)})$ **then**
    $F^{(t+1)} = \check{F}, \quad W^{(t+1)} = \check{W}$
    $s_1^{(t)} = F^{(t+1)} - F^{(t)}, \quad s_2^{(t)} = W^{(t+1)} - W^{(t)}$
    $y_1^{(t)} = \nabla g_F(F^{(t+1)}, W^{(t+1)}) - \nabla g_F(F^{(t)}, W^{(t)})$
    $y_2^{(t)} = \nabla g_W(F^{(t+1)}, W^{(t+1)}) - \nabla g_W(F^{(t)}, W^{(t)})$
    Compute $b_t = \langle s_1^{(t)}, y_1^{(t)}\rangle + \langle s_2^{(t)}, y_2^{(t)}\rangle$
    if $b_t \leqslant 0$ **then** $\alpha_{t+1}^+ = \alpha_{max}^+$
    else
      Compute $a_t = \langle s_1^{(t)}, s_1^{(t)}\rangle + \langle s_2^{(t)}, s_2^{(t)}\rangle$
      Compute $\alpha_{t+1}^+ = \min\{\alpha_{max}^+, \max\{\alpha_{min}^+, \frac{a_t}{b_t}\}\}$
    end if
    $t = t + 1$
  else
    $\lambda_{new} \in [\delta_1\lambda, \ \delta_2\lambda]$
    $\lambda = \lambda_{new}$
  end if
until *Convergence according to (19)* ▷ KKT conditions
Return W^*

IV. Experiments

To validate our method for action recognition in videos, we first demonstrate Fisher vector (FV) used for data representation. Then, we conduct extensive experiments on challenging data sets to test our framework's performance.

TABLE I

COMPARISON WITH IDT-BASED HAND-CRAFTED FEATURES (AVERAGE ACCURACY ± STD) WHEN $3 \times c$ TRAINING VIDEOS ARE LABELED

	JHMDB	HMDB51	UCF50	UCF101
Ours	**0.4238 ± 0.0185**	**0.2738 ± 0.0115**	**0.6184 ± 0.0174**	**0.5075 ± 0.0132**
SFUS	0.3258 ± 0.0243	0.1973 ± 0.0147	0.5465 ± 0.0177	0.4315 ± 0.0168
SFCM	0.3440 ± 0.0187	0.2246 ± 0.0120	0.5571 ± 0.0166	0.4362 ± 0.0150
MFCU	0.3552 ± 0.0167	0.2405 ± 0.0127	0.5803 ± 0.0185	0.4588 ± 0.0133
SVM-χ^2	0.3324 ± 0.0213	0.2019 ± 0.0132	0.5420 ± 0.0198	0.4204 ± 0.0162
SVM-linear	0.3773 ± 0.0180	0.2351 ± 0.0168	0.5851 ± 0.0182	0.4681 ± 0.0145

TABLE II

COMPARISON WITH IDT-BASED HAND-CRAFTED FEATURES (AVERAGE ACCURACY ± STD) WHEN $5 \times c$ TRAINING VIDEOS ARE LABELED

	JHMDB	HMDB51	UCF50	UCF101
Ours	**0.4658 ± 0.0178**	**0.3293 ± 0.0118**	**0.6743 ± 0.0187**	**0.5963 ± 0.0113**
SFUS	0.3775 ± 0.0189	0.2645 ± 0.0083	0.5905 ± 0.0183	0.5265 ± 0.0121
SFCM	0.3998 ± 0.0233	0.2719 ± 0.0069	0.6183 ± 0.0240	0.5264 ± 0.0135
MFCU	0.4135 ± 0.0152	0.2830 ± 0.0122	0.6373 ± 0.0177	0.5486 ± 0.0128
SVM-χ^2	0.3748 ± 0.0143	0.2616 ± 0.0153	0.6007 ± 0.0164	0.5127 ± 0.0157
SVM-linear	0.4120 ± 0.0122	0.2902 ± 0.0058	0.6320 ± 0.0212	0.5501 ± 0.0140

TABLE III

COMPARISON WITH IDT-BASED HAND-CRAFTED FEATURES (AVERAGE ACCURACY ± STD) WHEN $10 \times c$ TRAINING VIDEOS ARE LABELED

	JHMDB	HMDB51	UCF50	UCF101
Ours	**0.5479 ± 0.0181**	**0.3980 ± 0.0073**	**0.8007 ± 0.0193**	**0.6843 ± 0.0128**
SFUS	0.4836 ± 0.0185	0.3177 ± 0.0126	0.7153 ± 0.0217	0.6280 ± 0.0135
SFCM	0.5029 ± 0.0169	0.3473 ± 0.0105	0.7311 ± 0.0180	0.6273 ± 0.0142
MFCU	0.5143 ± 0.0173	0.3598 ± 0.0082	0.7558 ± 0.0178	0.6396 ± 0.0078
SVM-χ^2	0.4618 ± 0.0226	0.3235 ± 0.0114	0.7384 ± 0.0207	0.6057 ± 0.0137
SVM-linear	0.4957 ± 0.0218	0.3617 ± 0.0078	0.7570 ± 0.0226	0.6425 ± 0.0153

TABLE IV

COMPARISON WITH IDT-BASED HAND-CRAFTED FEATURES (AVERAGE ACCURACY ± STD) WHEN $15 \times c$ TRAINING VIDEOS ARE LABELED

	JHMDB	HMDB51	UCF50	UCF101
Ours	**0.5822 ± 0.0093**	**0.4407 ± 0.0092**	**0.8578 ± 0.0085**	**0.7218 ± 0.0076**
SFUS	0.5341 ± 0.0120	0.3613 ± 0.0131	0.7811 ± 0.0091	0.6437 ± 0.0122
SFCM	0.5278 ± 0.0118	0.3822 ± 0.0099	0.7955 ± 0.0089	0.6320 ± 0.0103
MFCU	0.5334 ± 0.0084	0.4035 ± 0.0077	0.8104 ± 0.0135	0.6701 ± 0.0078
SVM-χ^2	0.5090 ± 0.0136	0.3637 ± 0.0125	0.7782 ± 0.0075	0.6324 ± 0.0119
SVM-linear	0.5341 ± 0.0085	0.4020 ± 0.0071	0.8104 ± 0.0083	0.6796 ± 0.0080

TABLE V

COMPARISON WITH CNN-BASED DEEP-LEARNED FEATURES (AVERAGE ACCURACY ± STD) WHEN $3 \times c$ TRAINING VIDEOS ARE LABELED

	JHMDB	HMDB51	UCF50	UCF101
Ours	**0.5020 ± 0.0165**	**0.3231 ± 0.0120**	**0.6829 ± 0.0174**	**0.6861 ± 0.0128**
SFUS	0.4309 ± 0.0134	0.2617 ± 0.0133	0.6208 ± 0.0177	0.6257 ± 0.0136
SFCM	0.4721 ± 0.0178	0.3011 ± 0.0108	0.6394 ± 0.0166	0.6429 ± 0.0132
MFCU	0.4783 ± 0.0153	0.3031 ± 0.0127	0.6543 ± 0.0185	0.6527 ± 0.0137
SVM-χ^2	0.4289 ± 0.0202	0.2608 ± 0.0112	0.6117 ± 0.0198	0.6231 ± 0.0121
SVM-linear	0.4534 ± 0.0180	0.2913 ± 0.0141	0.6245 ± 0.0182	0.6447 ± 0.0133

TABLE VI

COMPARISON WITH CNN-BASED DEEP-LEARNED FEATURES (AVERAGE ACCURACY ± STD) WHEN $5 \times c$ TRAINING VIDEOS ARE LABELED

	JHMDB	HMDB51	UCF50	UCF101
Ours	**0.6066 ± 0.0167**	**0.3954 ± 0.0088**	**0.7646 ± 0.0193**	**0.7390 ± 0.0111**
SFUS	0.5219 ± 0.0181	0.3296 ± 0.0110	0.7323 ± 0.0217	0.6841 ± 0.0105
SFCM	0.5591 ± 0.0160	0.3540 ± 0.0092	0.7472 ± 0.0180	0.7098 ± 0.0113
MFCU	0.5667 ± 0.0153	0.3723 ± 0.0101	0.7472 ± 0.0178	0.7145 ± 0.0083
SVM-χ^2	0.5231 ± 0.0201	0.3257 ± 0.0123	0.7281 ± 0.0207	0.6892 ± 0.0127
SVM-linear	0.5579 ± 0.0174	0.3551 ± 0.0097	0.7323 ± 0.0226	0.7187 ± 0.0126

TABLE VII

COMPARISON WITH CNN-BASED DEEP-LEARNED FEATURES (AVERAGE ACCURACY ± STD) WHEN $10 \times c$ TRAINING VIDEOS ARE LABELED

	JHMDB	HMDB51	UCF50	UCF101
Ours	**0.7284 ± 0.0157**	**0.4897 ± 0.0106**	**0.8427 ± 0.0187**	**0.8477 ± 0.0112**
SFUS	0.6723 ± 0.0164	0.4172 ± 0.0112	0.7844 ± 0.0183	0.8054 ± 0.0101
SFCM	0.6934 ± 0.0183	0.4423 ± 0.0087	0.7993 ± 0.0240	0.8107 ± 0.0114
MFCU	0.7034 ± 0.0145	0.4623 ± 0.0134	0.8141 ± 0.0177	0.8266 ± 0.0115
SVM-χ^2	0.6710 ± 0.0134	0.4206 ± 0.0152	0.7767 ± 0.0164	0.8001 ± 0.0137
SVM-linear	0.6909 ± 0.0132	0.4512 ± 0.0057	0.7770 ± 0.0212	0.8173 ± 0.0103

TABLE VIII

COMPARISON WITH CNN-BASED DEEP-LEARNED FEATURES (AVERAGE ACCURACY ± STD) WHEN $15 \times c$ TRAINING VIDEOS ARE LABELED

	JHMDB	HMDB51	UCF50	UCF101
Ours	**0.7410 ± 0.0082**	**0.5830 ± 0.0090**	**0.8899 ± 0.0085**	**0.8683 ± 0.0078**
SFUS	0.6923 ± 0.0113	0.5200 ± 0.0123	0.8253 ± 0.0091	0.7898 ± 0.0102
SFCM	0.7110 ± 0.0100	0.5373 ± 0.0108	0.8290 ± 0.0089	0.8070 ± 0.0084
MFCU	0.7148 ± 0.0089	0.5542 ± 0.0087	0.8513 ± 0.0135	0.8419 ± 0.0087
SVM-χ^2	0.6941 ± 0.0116	0.5189 ± 0.0115	0.8179 ± 0.0075	0.8131 ± 0.0098
SVM-linear	0.7134 ± 0.0086	0.5370 ± 0.0068	0.8439 ± 0.0083	0.8448 ± 0.0070

A. Data Sets

In the experiments, three data sets are used, including the JHMDB data set [70], the HMDB51 data set [71], the UCF50 data set [72], and the UCF101 data set [73]. The **JHMDB** data set is a subset of HMDB51 with 928 clips comprising 21 action categories. The **HMDB51** data set contains 6766 video sequences recording 51 action categories. The **UCF50** data set has 50 action categories, consisting of real-world videos taken from YouTube. There are 6618 video clips in UCF50. The **UCF101** data set collects 13 320 video clips including 101 action categories. As far as the testing set, we use the standard testing set provided by the authors on JHMDB and HMDB51 data sets, and the testing set of the first split on UCF50 and UCF101 data sets. Due to the random training samples selection, we repeat the experiment for 10 trials to avoid any bias. The average accuracy and standard deviation are reported.

For the JHMDB and HMDB51 data sets, we follow [30] and [31] and use the standard data partition provided by the author. For the UCF50 and UCF101 data sets, unlike [30] and [31] that randomly split each data set into training and testing sets, we only use the first split provided by author due to computation complexity and limited memory resource. In addition, we randomly select 30 videos per category as the training data including the labeled and unlabeled samples and apply the original testing sets for comparison in a more fair way.

B. Features

For hand-crafted features, we extract improved dense trajectories (IDTs)-based features with HOG + HOF + MBH descriptors [74]. The dimension D is reduced to 198 by performing PCA and L2-normalization. After training a GMM codebook with K Gaussians based on 256 000 randomly sampled features, each action video is represented by a $2DK = 6336$ dimensional FV with Power L2-normalization, if $K = 16$ as TablesI–IV.

For deep-learned features, the convolutional neural networks (CNN)-based features are selected, e.g., the trajectory-pooled

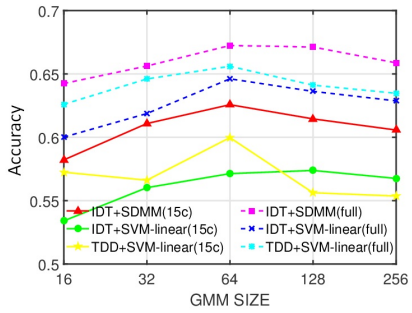


Fig. 3. Comparisons on JHMDB with respect to different gmmSize of FV encoding.

deep-convolutional descriptors (TDDs) [75] and temporal segment networks (TSNs) [76]. We follow [75] to concatenate eight normalized deep-learned features from spatial conv4 + conv5 and temporal conv3 + conv4 layers, let the dimension of combined TDDs becomes $D = 64 \times 8 = 512$, since each TDDs' dimension of a video is decorrelated to 64 by PCA. Then, we encode the combined TDDs into FV representation, and the final dimension of each video can be changed to $2DK = 16\,384$ when $K = 16$, as shown in Fig. 3. Meanwhile, the TSN models of $3 \times c$, $5 \times c$, $10 \times c$, and $15 \times c$ are retrained according to [76], then we extract the global pool features of $3 \times c$, $5 \times c$, $10 \times c$, and $15 \times c$, respectively, by corresponding trained TSN model, concatenate rgb + flow into 2048 dimension with Power L2-normalization, as shown in Tables V–VIII.

C. Experimental Setup

To evaluate the performance of our approach, the proposed algorithm is compared to the five state-of-the-art methods which include SVM with χ^2 kernel, SVM with linear kernel, SFCM [30], SFUS [4], and MFCU [31].

Note that SFCM, SFUS, and MFCU are semisupervised learning approaches. SFCM and MFCU also exploit the data manifold and are designed for action recognition. To demonstrate the superiority of our method, we employed these related state-of-the-art methods for comparison. Also, with the available source codes, we can run experiments on different data sets and settings to facilitate fair comparisons.

For training phase, we denote c as the class number for each data set ($c = 21$, 51, 50, and 101 for JHMDB, HMDB51, UCF50, and UCF101, respectively). As semisupervised training set contains both labeled and unlabeled data, we randomly select 30 videos per category in the training set, where m labeled videos ($m = 3$, 5, 10, and 15) per category are sampled, thus resulting in $3 \times c$, $5 \times c$, $10 \times c$, and $15 \times c$ randomly labeled videos, while the remaining training videos are unlabeled.

For testing phase, we use the standard testing set provided by the author on JHMDB and HMDB51 data sets, and the first split of the testing set on UCF50 and UCF101 data sets due to the limited memory resource.

TABLE IX
AVERAGE RUN TIMES (IN SECONDS) ON JHMDB

Ours	SFUS	SFCM	MFCU
41.93	44.63	173.93	104.83

For semisupervised parameters, including SFUS, SFCM, MFCU, and our SDMM's α, β, μ, we follow the same setting utilized in [30] and [31] and range from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10^1, 10^2, 10^3, 10^4\}$.

For SPG parameters, since they are not sensitive to our algorithm, we follow [38] and set $M = 10, \alpha_{\min} = 10^{-15}$, and $\alpha_{\max} = 10^{15}$, sufficient decrease parameter $\gamma = 10^{-4}$, safeguarding parameters $\delta_1 = 0.1, \delta_2 = 0.9$, and $\lambda_{\text{new}} = (1/2)(\delta_1\lambda + \delta_2\lambda)$. Initially, $\alpha_0 \in [\alpha_{\min}, \alpha_{\max}]$ is arbitrary, we set $\alpha_0 = 1$ in our experiments. Technically, since the dimension of FV is relatively high, it is hard to stop iteration for merely subtracting the last two objective function values, we regard the relative error of the objective function values as iteration stopping criterion in Algorithm 1. The nonnegative small constant ε of (19) is suggested to set 10^{-6}.

D. Comparison Results

Tables I–VIII show the action recognition results on four challenging data sets with respect to different number of labeled training data. Specifically, we compare the proposed method to those other approaches that only apply a single type of feature, i.e., FV representation.

1) Performance on Action Recognition: We observe the following.

1) Our method consistently obtains the best recognition performance, the recognition of our semisupervised classifiers even better than the popular supervised classifiers such as linear SVM.
2) We verify the effectiveness of the proposed method with IDT-and CNN-based representations beyond Bag-of-Words.
3) All methods achieve worse results on HMDB51 compared with those on another three data sets. This is probably owed to the complexity of HMDB51.
4) The recognition accuracy of all methods is improved with the increase of the number of labeled training videos.
5) Our method gains better performance when the amount of labeled data is small. For example, when only $3 \times c$ (63 out of 660 training data for JHMDB) training data are labeled, our method achieves the recognition accuracy of 42.38%, which is better than others.

These results indicate that our algorithm benefits from the multimanifold analysis of feature correlations.

The fully supervised linear SVM is taken as baseline, we average the accuracy of $3 \times c$, $5 \times c$, $10 \times c$, and $15 \times c$ cases totally. Using the IDT features, the average accuracy of our SDMM on JHMDB, HMDB51, UCF50, and

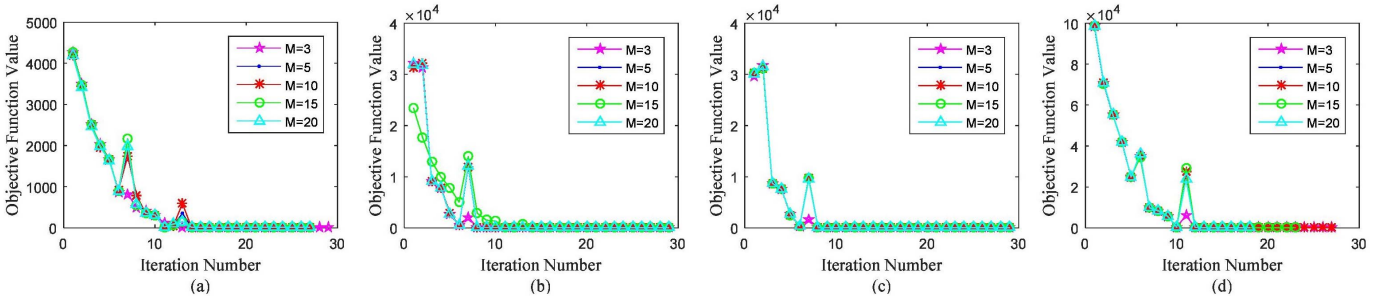


Fig. 4. Convergence curves of the objective function values in (15) by using our SDMM on four data sets. Objective function values become optimum solution by applying the proposed algorithm with KKT conditions. (a) JHMDB. (b) HMDB51. (c) UCF50. (d) UCF101.

UCF101 is improved by 5.02%, 3.82%, 4.16%, and 4.24%, respectively. While using the TSN features, the average accuracy of our SDMM on JHMDB, HMDB51, UCF50, and UCF101 is improved by 4.06%, 3.92%, 5.06%, and 3.39%, respectively, as compared with linear SVM. It is evident that by leveraging the unlabeled data, the recognition performance can be improved.

The deep learning approaches, which trained on large-scale labeled data, have shown promising performance on image classification and action recognition. To validate the performance of the deep learning approaches on small-scale data set, we follow the experimental setup of [75] to extract TDDs, then encode the TDDs of every video into FV representation and recognize actions by linear SVM on JHMDB. In order to train GMMs with K ($K = 256$), we decorrelate TDDs with PCA and reduce its dimension to $D = 64$. Note that we utilize the combined TDDs from spatial conv4 + conv5 and temporal conv3 + conv4 nets.

We compare our SDMM algorithm with linear SVM and TDD in case of $15 \times c$, the comparison results on JHMDB data set with respect to different gmmSize is shown in Fig. 3. Note that this figure contains both semisupervised learning and fully supervised learning which use $15 \times c$ case of training sets and full set of training set, respectively. As we expect, either semisupervised learning or fully supervised learning, the performance of SDMM consistently better than linear SVM and TDD.

These results may account for many reasons. First, our method not only takes the advantage of compared semisupervised approaches in [4], [30], [31], and [37] but also leverages the intraclass compactness and interclass separability simultaneously, hence our performance gain over other methods is more significant when the labeled data are small. Second, we enlarge the geometric structure information of feature subspace by increasing training samples with many unlabeled samples for discriminant learning, and the objective function optimization is solved by the SPG method and the KKT conditions mathematically, thus our multimanifold works well in the small labeled data case. At last, the deep learning approaches that are trained such as TDD built on CNN with deep layers, the spatial net, and the temporal net rely on large-scale samples. However, small-scale data set such as real-world surveillance applications, which are hard to collect labeled video data from daily life, cannot adapt to deep learning approaches, because the scale of network weights, which are learned by using fine-tuned network structure based on large-scale data sets, may be larger than the scale of action features.

2) Convergence Study: To validate the proposed algorithm that it can derive optimum solution by the SPG method and the KKT conditions, we conduct experiments on all four data sets by applying convergence curves of the objective function values. The number of labeled training samples is set to $15 \times c$ for each data set, and the parameters are set to the median value of the tuned range. The results in Fig. 4 demonstrate that the objective function values converge after only a few iterations. Note that there are oscillations caused by the SPG method in our convergence curves, the objective values are not monotonically decreasing before iterations stop.

3) Computation Speed: We also set a practical example for comparing the computation speed of the aforementioned semisupervised algorithms. We consider the case of $15 \times c$ labeled samples for JHMDB, and use the training–testing set given in Section IV-C, train GMMs with $K = 16$ and then compute the average run time of algorithms over the standard splits. Given the high dimension of raw features are utilized in SFUS, SFCM, and MFCU, we first perform SVD to reduce raw features' dimension according to [37]. Nevertheless, our SDMM still obtains the fastest speed due to the trait of the SPG method. Compared with the SFUS, SFCM, and MFCU, the run time of SDMM gains $1.06\times$, $4.15\times$, and $2.50\times$ faster, respectively, as shown in Table IX.

4) Parameter Sensitivity Study: Our algorithm involves two types of parameters, i.e., semisupervised parameters and SPG parameters. To learn how they affect the analysis performance and iteration process on action recognition, we conduct extensive experiments on the parameter sensitivity.

For semisupervised parameters, we first verify that SDMM benefits from intramanifold and intermanifold by multimanifold discriminant analysis in Fig. 5(a) and (b). The JHMDB and HMDB51 data sets are taken to study the impact of multimanifold learning. We fix α and μ at their optimal values over the second split, i.e., 10^{-3} and 10^{3}, respectively, for $15 \times c$ labeled training data. It can be seen that as β varies from 10^{-4} to 10^{-2}, the accuracy increases accordingly and reaches to the peak value when $\beta = 10^{-2}$. Note that Fig. 5(a) and (b) can be regarded as the influence of both intramanifold and intermanifold structure proportion on accuracy.

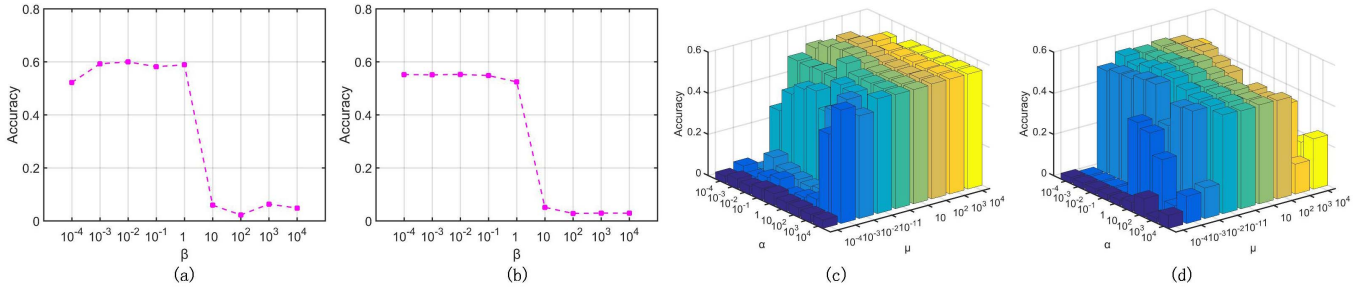


Fig. 5. (a) The variation of accuracy on JHMDB using IDT, w.r.t. the parameter β with fixed α and μ. (b) The variation of accuracy on HMDB51 using TSN, w.r.t. the parameter β with fixed α and μ. (c) The variation of accuracy on JHMDB using IDT, w.r.t. different α, μ while $\beta = 10^{-2}$. (d) The variation of accuracy on HMDB51 using TSN, w.r.t. different α, μ while $\beta = 10^{-2}$.

Since we perceive the proportion of intramanifold structure as constant 1, hence a larger $\frac{\beta}{1}$ means a larger proportion of intermanifold structural consideration, and vice versa. When $\beta = 0$, no intermanifold structure is utilized, thus, if $\beta \to +\infty$, no intramanifold structure is contained. The results illustrate that appropriately exploiting intraclass compactness and interclass separability simultaneously in multimanifold subspace can further improve the performance. Then, we keep $\beta = 10^{-2}$, and show the parameter sensitivity results in Fig. 5(c) and (d). From these figures, we can see that mining correlations between multiple related tasks are beneficial to improve the performance. More specifically, we conduct extensive experiments on HMDB51 using TSN features, as shown in Fig. 5(b) and (d). Fig. 5 shows that the recognition can achieve stable high accuracy when all the hyperparameters are selected in certain range, e.g., α ranges in $\{10^{-2}, 10^{-1}, 1\}$, β ranges in $\{10^{-3}, 10^{-2}, 10^{-1}\}$, and μ ranges in $\{10^{-1}, 1, 10^{1}\}$. In other words, there is flexibility in choosing the parameters in order to achieve optimal performance.

For SPG parameters, accuracy and M are used to reflect the performance and iteration variation, respectively, where M denotes the number of former iteration which is designed for inequality calculation. In algorithm 1, the step 6 of SPG method, new objective function value $g(\check{F}, \check{W})$ is supposed to compare with the former Mth objective function values. Fig. 4 illustrates the iteration variation with respect to M on four databases. In Fig. 4, the iteration process changes slightly corresponding to different values of M. The impact of different values of these parameters is supposed to be related to the trait of the feature representation. Generally speaking, M is not sensitive to the iteration of SDMM.

V. Conclusion

In this paper, a novel algorithm is proposed to categorize human actions in videos by exploring data distribution and feature correlation. Using a multimanifold-based joint framework, our method discovers the intrinsic relationship of midlevel features to improve recognition performance. Second, the SPG method and the KKT conditions are applied to optimize the objective function for training robust classifiers. Finally, we extend the classifier into the semisupervised scenario to exploit both labeled and unlabeled videos. We evaluate our framework for action recognition on four challenging data sets. The experimental results show that our approach outperforms all compared algorithms, especially when the amount of labeled data is relatively small. Since semisupervised learning methods based on generative adversarial networks (GANs) have obtained strong empirical results, we prepare to discover discriminative information via GANs with shallow layers in the future.

References

[1] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

[2] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 661–669, Apr. 2013.

[3] X. Chang and Y. Yang, "Semisupervised feature analysis by mining correlations among multiple tasks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2294–2305, Oct. 2017.

[4] Z. Ma, F. Nie, Y. Yang, J. R. R. Uijlings, and N. Sebe, "Web image annotation via subspace-sparsity collaborated feature selection," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1021–1030, Aug. 2012.

[5] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[6] Z. Xu *et al.*, "Action recognition by saliency-based dense sampling," *Neurocomputing*, vol. 236, pp. 82–92, May 2016.

[7] W. Zuo, D. Ren, S. Gu, L. Lin, and L. Zhang, "Discriminative learning of iteration-wise priors for blind deconvolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3232–3240.

[8] W. Zuo, L. Zhang, C. Song, D. Zhang, and H. Gao, "Gradient histogram estimation and preservation for texture enhanced image denoising," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2459–2472, Jun. 2014.

[9] J. Ma, J. Zhao, Y. Ma, and J. Tian, "Non-rigid visible and infrared face registration via regularized Gaussian fields criterion," *Pattern Recognit.*, vol. 48, no. 3, pp. 772–784, 2015.

[10] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4792–4800.

[11] Z. Wang, N. M. Nasrabadi, and T. S. Huang, "Semisupervised hyperspectral classification using task-driven dictionary learning with laplacian regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1161–1173, Mar. 2015.

[12] Z. Wang, H. Li, Q. Ling, and W. Li, "Robust temporal-spatial decomposition and its applications in video processing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 3, pp. 387–400, Mar. 2013.

[13] X. Peng, M. Yuan, Z. Yu, Y. Y. Wei, and L. Zhang, "Semi-supervised subspace learning with L2graph," *Neurocomputing*, vol. 208, pp. 143–152, Oct. 2016.

[14] X. Peng, C. Lu, Y. Zhang, and H. Tang, "Connections between nuclear-norm and Frobenius-norm-based representations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 218–224, Jan. 2018.

[15] X. Peng, L. Zhang, and Z. Yi, "Scalable sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 430–437.

[16] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015.

[17] L. Shao, L. Liu, and X. Li, "Feature learning for image classification via multiobjective genetic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1359–1371, Jul. 2014.

[18] L. Shao, L. Liu, and M. Yu, "Kernelized multiview projection for robust action recognition," *Int. J. Comput. Vis.*, vol. 118, no. 2, pp. 115–129, 2016.

[19] J. Qin *et al.*, "Zero-shot action recognition with error-correcting output codes," in *Proc. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1042–1051.

[20] M. Yu, L. Liu, and L. Shao, "Structure-preserving binary representations for rgb-d action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1651–1664, Aug. 2016.

[21] D. V. Prokhorov *et al.*, "IEEE transactions on intelligent vehicles senior associate editors," *IEEE Trans. Intell. Veh.*, vol. 1, no. 1, pp. 3–5, Mar. 2016.

[22] A. Wendel, M. Maurer, G. Graber, T. Pock, and H. Bischof, "Dense reconstruction on-the-fly," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1450–1457.

[23] Z. Shao, J. Cai, and Z. Wang, "Smart monitoring cameras driven intelligent processing to big surveillance video data," *IEEE Trans. Big Data*, vol. 4, no. 1, pp. 105–116, Mar. 2018.

[24] Z. Wang *et al.*, "Zero-shot person re-identification via cross-view consistency," *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 260–272, Feb. 2016.

[25] Z. Wang *et al.*, "Person reidentification via discrepancy matrix and matrix metric," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 3006–3020, Oct. 2018.

[26] R. Lan and Y. Zhou, "Quaternion-Michelson descriptor for color image classification," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5281–5292, Nov. 2016.

[27] R. Lan, Y. Zhou, and Y. Y. Tang, "Quaternionic Weber local descriptor of color images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 2, pp. 261–274, Feb. 2017.

[28] R. Hou, C. Chen, and M. Shah, "Tube convolutional neural network (T-CNN) for action detection in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1–10.

[29] Z. Shao, W. Wu, Z. Wang, W. Du, and C. Li, "SeaShips: A large-scale precisely annotated dataset for ship detection," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2593–2604, Oct. 2018.

[30] S. Wang, Y. Yang, Z. Ma, X. Li, C. Pang, and A. G. Hauptmann, "Action recognition by exploring data distribution and feature correlation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1370–1377.

[31] S. Wang, Z. Ma, Y. Yang, X. Li, C. Pang, and A. G. Hauptmann, "Semi-supervised multiple feature analysis for action recognition," *IEEE Trans. Multimedia*, vol. 16, no. 2, pp. 289–298, Feb. 2014.

[32] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *Proc. IEEE 11th Int. Conf. Comput. Vis. (ICCV)*, Oct. 2007, pp. 1–7.

[33] D. Cai and X. He, "Manifold adaptive experimental design for text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 4, pp. 707–719, Apr. 2012.

[34] J. Jiang, R. Hu, Z. Han, and K. Huang, "Graph discriminant analysis on multi-manifold (GDAMM): A novel super-resolution method for face recognition," in *Proc. IEEE Int. Conf. Image Process.*, Sep./Oct. 2012, pp. 1465–1468.

[35] J. Jiang, R. Hu, Z. Han, L. Chen, and J. Chen, "Coupled discriminant multi-manifold analysis with application to low-resolution face recognition," in *MultiMedia Modeling*, vol. 8935. Cham, Switzerland: Springer, 2015, pp. 37–48.

[36] J. Jiang, R. Hu, Z. Wang, and Z. Cai, "CDMMA: Coupled discriminant multi-manifold analysis for matching low-resolution face images," *Signal Process.*, vol. 124, pp. 162–172, 2016.

[37] Y. Yang, F. Wu, F. Nie, H. T. Shen, Y. Zhuang, and A. G. Hauptmann, "Web and personal image annotation by mining label correlation with relaxed visual graph embedding," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1339–1351, Mar. 2012.

[38] A. Bouhamidi, K. Jbilou, and M. Raydan, "Convex constrained optimization for large-scale generalized Sylvester equations," *Comput. Optim. Appl.*, vol. 48, no. 2, pp. 233–253, Mar. 2011.

[39] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 38–52, Feb. 2011.

[40] E. G. Birgin, J. M. Martínez, and M. Raydan, "Nonmonotone spectral projected gradient methods on convex sets," *SIAM J. Optim.*, vol. 10, no. 4, pp. 1196–1211, 2000.

[41] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 10, pp. 2756–2779, 2007.

[42] J.-F. Li and Z.-Y. Peng, "A hybrid algorithm for solving minimization problem over (*R,S*)-symmetric matrices with the matrix inequality constraint," *Linear Multilinear Algebra*, vol. 63, no. 5, pp. 1049–1072, 2014.

[43] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[44] S. Abu-El-Haija *et al.* (2016). "YouTube-8M: A large-scale video classification benchmark." [Online]. Available: https://arxiv.org/abs/1609.08675

[45] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 961–970.

[46] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jan. 2018, pp. 1–10.

[47] D. George *et al.*, "A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs," *Science*, vol. 358, no. 6368, p. eaag2612, 2017.

[48] M. Luo, X. Chang, L. Nie, Y. Yang, A. G. Hauptmann, and Q. Zheng, "An adaptive semisupervised feature analysis for video semantic recognition," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 648–660, Feb. 2018.

[49] Y. Yan, C. Xu, D. Cai, and J. J. Corso, "Weakly supervised actor-action segmentation via robust multi-task ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1022–1031.

[50] W. Yu and C. Zhao, "Online fault diagnosis in industrial processes using multimodel exponential discriminant analysis algorithm," *IEEE Trans. Control Syst. Technol.*, to be published.

[51] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Lovell, "Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching," in *Proc. CVPR*, Jun. 2011, pp. 2705–2712.

[52] P. Li, J. Bu, C. Chen, Z. He, and D. Cai, "Relational multimanifold coclustering," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1871–1881, Dec. 2013.

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

[54] W. Yu and C. Zhao, "Sparse exponential discriminant analysis and its application to fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5931–5940, Jul. 2018.

[55] J. Ma, J. Wu, J. Zhao, J. Jiang, H. Zhou, and Q. Z. Sheng, "Nonrigid point set registration with robust transformation learning under manifold regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2018.2872528.

[56] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Face super-resolution via multilayer locality-constrained iterative neighbor embedding and intermediate dictionary learning," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4220–4231, Oct. 2014.

[57] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Noise robust face hallucination via locality-constrained representation," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1268–1281, Aug. 2014.

[58] J. Jiang, J. Ma, C. Chen, X. Jiang, and Z. Wang, "Noise robust face image super-resolution through smooth sparse representation," *IEEE Trans. Cybern.*, vol. 47, no. 11, pp. 3991–4002, Nov. 2017.

[59] F. Nie, D. Xu, I. W. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Trans. Image Process.*, vol. 19, no. 7, pp. 1921–1932, Jul. 2010.

[60] Y. Gao, J. Ma, and A. L. Yuille, "Semi-supervised sparse representation based classification for face recognition with insufficient labeled samples," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2545–2560, May 2017.

[61] J. Ma, J. Jiang, C. Liu, and Y. Li, "Feature guided Gaussian mixture model with semi-supervised EM and local geometric constraint for retinal image registration," *Inf. Sci.*, vol. 417, pp. 128–142, Nov. 2017.

[62] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.

[63] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *Int. J. Comput. Vis.*, vol. 113, no. 2, pp. 113–127, 2015.
[64] Z. Ma, Y. Yang, F. Nie, N. Sebe, S. Yan, and A. G. Hauptmann, "Harnessing lab knowledge for real-world action recognition," *Int. J. Comput. Vis.*, vol. 109, nos. 1–2, pp. 60–73, 2014.
[65] Y. Yang, Z. Ma, Y. Yang, F. Nie, and H. T. Shen, "Multitask spectral clustering by exploring intertask correlation," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 1083–1094, May 2015.
[66] G. M. Fung and O. L. Mangasarian, "Multicategory proximal support vector machine classifiers," *Mach. Learn.*, vol. 59, nos. 1–2, pp. 77–97, 2005.
[67] R. Wang and X. Chen, "Manifold discriminant analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 429–436.
[68] F. Chung, *Spectral Graph Theory*. Providence, RI, USA: AMS, 1997.
[69] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16. 2004, pp. 321–328.
[70] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3192–3199.
[71] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 2556–2563.
[72] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of Web videos," *Mach. Vis. Appl.*, vol. 24, no. 5, pp. 971–981, 2013.
[73] K. Soomro, A. R. Zamir, and M. Shah. (2012). "UCF101: A dataset of 101 human actions classes from videos in the wild." [Online]. Available: https://arxiv.org/abs/1212.0402
[74] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *Int. J. Comput. Vis.*, vol. 119, no. 3, pp. 219–238, 2016.
[75] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4305–4314.
[76] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.*, vol. 9912. Amsterdam, The Netherlands: Springer, 2016, pp. 20–36.

Zengmin Xu received the B.S. and M.S. degrees from the Changsha University of Science and Technology, Changsha, China, in 2003 and 2006, respectively. He is currently pursuing the Ph.D. degree with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan, China.

His current research interests include multimedia content analysis, computer vision, and pattern recognition.

Ruimin Hu (SM'10) received the B.S. and M.S. degrees from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1984 and 1990, and Ph.D. degree in communication and electronic system from the Huazhong University of Science and Technology, Wuhan, China, in 1994.

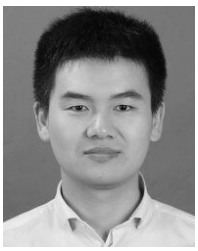
He was the Executive Chairman of the Audio Video Coding Standard Workgroup of China, Beijing, China, in audio section. He is currently the Director of the National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan, China, and also with the Hubei Key Laboratory of Multimedia Network Communication Engineering, Wuhan University. He has published two books and over 300 scientific papers. His current research interests include audio/video coding and decoding, video surveillance, and multimedia data processing.

Jun Chen received the M.S. degree in instrumentation from the Huazhong University of Science and Technology, Wuhan, China, in 1997, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, in 2008.

He is currently the Deputy Director of the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University. His current research interests include multimedia communications and security emergency information processing.

Chen Chen received the B.S. degree in automation from Beijing Forestry University, Beijing, China, in 2009, the M.S. degree in electrical engineering from Mississippi State University, Starkville, MS, USA, in 2012, and the Ph.D. degree from the Department of Electrical Engineering, The University of Texas at Dallas, Richardson, TX, USA, in 2016.

He was a Post-Doctoral Research Associate with the Center for Research in Computer Vision, University of Central Florida, Orlando, FL, USA, from 2016 to 2018. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, Charlotte, NC, USA. His current research interests include signal and image processing, computer vision, and deep learning. He published over 50 papers in refereed journals and conferences in these areas.



Junjun Jiang received the B.S. degree from the Department of Mathematics, Huaqiao University, Quanzhou, China, in 2009, and the Ph.D. degree from the School of Computer Science, Wuhan University, Wuhan, China, in 2014.

From 2015 to 2018, he was an Associate Professor with the China University of Geosciences, Wuhan. Since 2016, he has been a Project Researcher with the National Institute of Informatics, Tokyo, Japan. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. His current research interests include image processing and computer vision.

Jiaofen Li received the B.S. degree from the Changsha University of Science and Technology, Changsha, China, in 2005, and the M.S. and Ph.D. degrees from the College of Mathematics and Econometrics, Hunan University, Changsha, in 2007 and 2010, respectively.

He is currently a Professor with the School of Mathematics and Computational Science, Guilin University of Electronic Technology, Guilin, China. His current research interests include inverse eigenvalue problem, linear matrix equation, and matrix algebra in control theory.

Hongyang Li received the M.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2005. He is currently pursuing the Ph.D. degree with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan.

His current research interests include computer vision and pattern recognition.